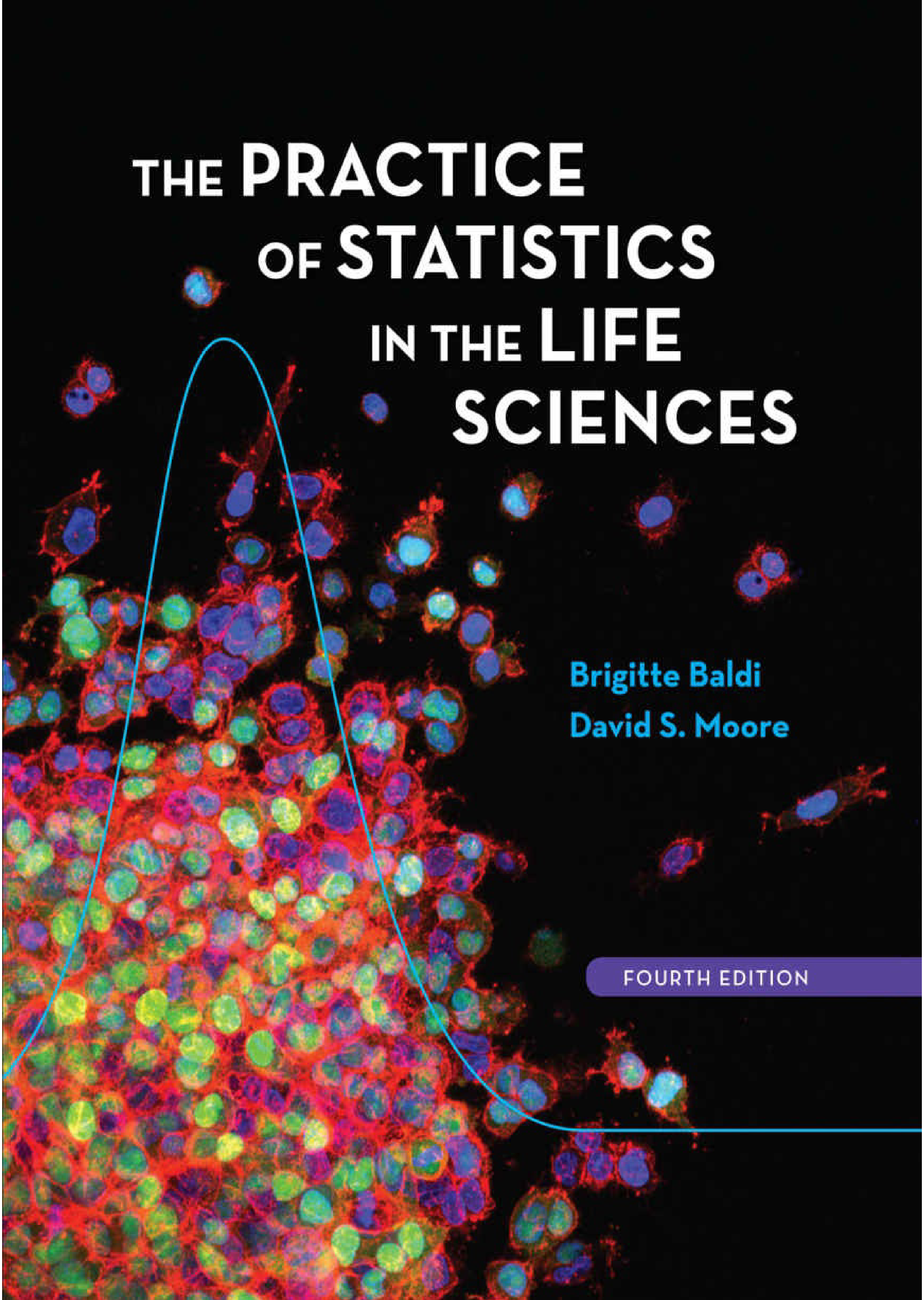
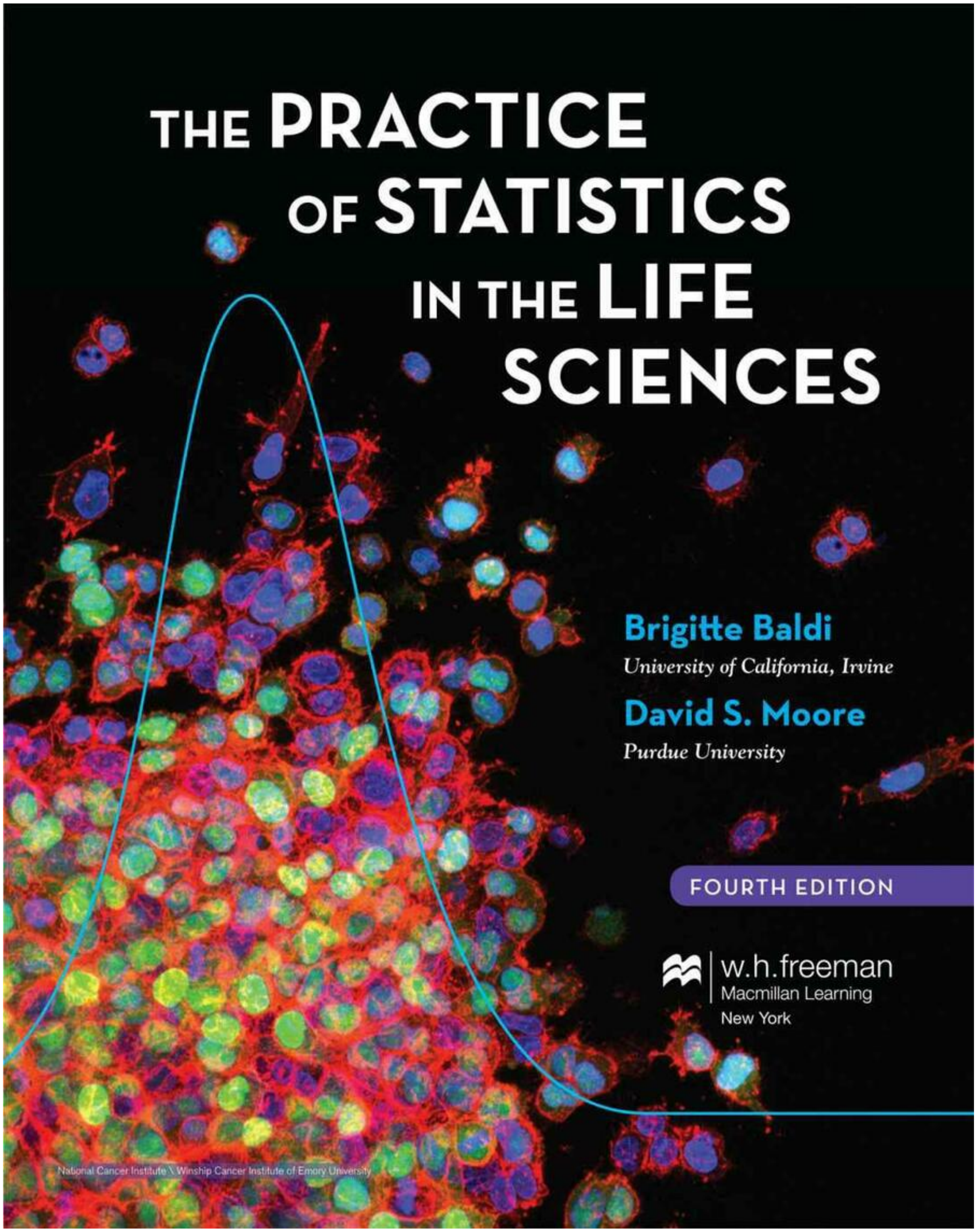


THE PRACTICE OF STATISTICS IN THE LIFE SCIENCES

Brigitte Baldi
David S. Moore

FOURTH EDITION



THE PRACTICE OF STATISTICS IN THE LIFE SCIENCES

Brigitte Baldi
University of California, Irvine
David S. Moore
Purdue University

FOURTH EDITION

w.h.freeman
Macmillan Learning
New York

National Cancer Institute \ Winship Cancer Institute of Emory University

Vice President, STEM: Ben Roberts
Senior Program Manager: Karen Carson
Marketing Manager: Tom DeMarco
Marketing Assistant: Savannah DiMarco
Executive Development Editor: Katrina Mangold
Executive Media Editor: Catriona Kaplan
Associate Editor: Andy Newton
Editorial Assistant: Katharine Munz
Director of Content Management Enhancement: Tracey Kuehn
Managing Editor: Lisa Kinne
Director of Design, Content Management: Diana Blume
Senior Design Manager: Vicki Tomaselli
Cover Designer: Vicki Tomaselli
Text Designer: Amanda Kavanaugh
Director of Digital Production: Keri deManigold
Senior Media Project Manager: Alison Lorber
Associate Media Project Manager: Hanna Squire
Senior Workflow Project Manager: Paul Rohloff
Senior Photo Editor: Robin Fadool
Photo Researcher: Naomi Kornhauser
Composition: Lumina Datamatics, Inc.
Cover Image: Front Cover Photo Credit: National Cancer Institute \
Winship Cancer Institute of Emory University
Back Cover Photo Credits:
Top: mark@rocketclips.com/Depositphotos.com;
middle: Science Source; bottom: encrier/Depositphotos.com

Library of Congress Control Number: 2017945523

ISBN 978-1-319-06749-6 (EPUB)

© 2018, 2014, 2012, 2009 by W. H. Freeman and Company
All rights reserved.

W. H. Freeman and Company
One New York Plaza
Suite 4500

New York, NY 10004-1562
www.macmillanlearning.com

Brief Contents

Part I **Collecting and Exploring Data**

EXPLORING DATA: VARIABLES AND DISTRIBUTIONS

Chapter 1 Picturing Distributions with Graphs

Chapter 2 Describing Quantitative Distributions with Numbers

EXPLORING DATA: RELATIONSHIPS

Chapter 3 Scatterplots and Correlation

Chapter 4 Regression

Chapter 5 Two-Way Tables[i]

PRODUCING DATA

Chapter 6 Samples and Observational Studies

Chapter 7 Designing Experiments

Chapter 8 Collecting and Exploring Data: Part I Review

Part II **From Chance to Inference**

PROBABILITY RULES

Chapter 9 Essential Probability Rules

Chapter 10 Independence and Conditional Probabilities[i]

PROBABILITY DISTRIBUTIONS

Chapter 11 The Normal Distributions

Chapter 12 Discrete Probability Distributions[i]

Chapter 13 Sampling Distributions

THE IDEA OF INFERENCE

Chapter 14 Introduction to Inference

Chapter 15 Inference in Practice

Chapter 16 From Chance to Inference: Part II Review

Part III Statistical Inference

INFERENCE ABOUT QUANTITATIVE VARIABLES

Chapter 17 Inference about a Population Mean

Chapter 18 Comparing Two Means

INFERENCE ABOUT CATEGORICAL VARIABLES

Chapter 19 Inference about a Population Proportion

Chapter 20 Comparing Two Proportions

Chapter 21 The Chi-Square Test for Goodness of Fit

INFERENCE ABOUT RELATIONSHIPS

Chapter 22 The Chi-Square Test for Two-Way Tables

Chapter 23 Inference for Regression

Chapter 24 One-Way Analysis of Variance: Comparing Several Means

Chapter 25 Statistical Inference: Part III Review

Part IV Optional Companion Chapters (available on the *PSLS 4e* Companion Website)

Chapter 26 More about Analysis of Variance: Follow-up Tests and Two-Way ANOVA

Chapter 27 Nonparametric Tests

Chapter 28 Multiple and Logistic Regression

Contents

To the Instructor: About this Book

Acknowledgments

Media and Supplements

To the Student: Statistical Thinking

About the Authors

Part I Collecting and Exploring Data

Chapter 1 Picturing Distributions with Graphs

Individuals and variables

Identifying categorical and quantitative variables

Categorical variables: pie charts and bar graphs

Quantitative variables: histograms

Interpreting histograms

Quantitative variables: dotplots

Time plots

Discussion: (Mis)adventures in data entry

Chapter 2 Describing Quantitative Distributions with Numbers

Measures of center: median, mean

Measures of spread: percentiles, standard deviation

Graphical displays of numerical summaries

Spotting suspected outliers[i]

Discussion: Dealing with outliers

Organizing a statistical problem

Chapter 3 Scatterplots and Correlation

Explanatory and response variables

Relationship between two quantitative variables: scatterplots

Adding categorical variables to scatterplots

Measuring linear association: correlation

Chapter 4 Regression

The least-squares regression line

Facts about least-squares regression

Outliers and influential observations

Working with logarithm transformations[i]

Cautions about correlation and regression

Association does not imply causation

Chapter 5 Two-Way Tables[i]

Marginal distributions

Conditional distributions

Simpson’s paradox

Chapter 6 Samples and Observational Studies

Observation versus experiment

Sampling

Sampling designs

Sample surveys

Cohorts and case-control studies

Chapter 7 Designing Experiments

Designing experiments

Randomized comparative experiments

Common experimental designs

Cautions about experimentation

Ethics in experimentation

Discussion: The Tuskegee syphilis study

Chapter 8 Collecting and Exploring Data: Part I Review

Part I summary

Comprehensive review exercises

Large data sets exercises

Online data sources

EESEE case studies

Part II From Chance to Inference

Chapter 9 Essential Probability Rules

The idea of probability

Probability models

Probability rules

Discrete versus continuous probability models

Random variables

Risk and Odds[i]

Chapter 10 Independence and Conditional Probabilities[i]

Relationships among several events

Conditional probability

General probability rules

Tree diagrams

Bayes's theorem

Discussion: Making sense of conditional probabilities in diagnostic tests

Chapter 11 **The Normal Distributions**

Normal distributions

The 68–95–99.7 rule

The standard Normal distribution

Finding Normal probabilities

Finding percentiles

Using the standard Normal table[i]

Normal quantile plots[i]

Chapter 12 **Discrete Probability Distributions**[i]

The binomial setting and binomial distributions

Binomial probabilities

Binomial mean and standard deviation

The Normal approximation to binomial distributions

The Poisson distributions

Poisson probabilities

Chapter 13 Sampling Distributions

Parameters and statistics

Statistical estimation and sampling distributions

The sampling distribution of x-$\bar{x}$

The central limit theorem

The sampling distribution of p^$\hat{p}$

The law of large numbers[i]

Chapter 14 Introduction to Inference

Statistical estimation

Margin of error and confidence level

Confidence intervals for the mean µμ

Hypothesis testing

P-value and statistical significance

Tests for a population mean

Tests from confidence intervals

Chapter 15 Inference in Practice

Conditions for inference in practice

How confidence intervals behave

How hypothesis tests behave

Discussion: The scientific approach

Planning studies: selecting an appropriate sample size[i]

Chapter 16 From Chance to Inference: Part II Review

Part II summary

Comprehensive review exercises

Advanced topics (optional material)

Online data sources

EESEE case studies

Part III Statistical Inference

Chapter 17 Inference about a Population Mean

Conditions for inference

The *t* distributions

The one-sample *t* confidence interval

The one-sample *t* test

Matched pairs *t* procedures

Robustness of *t* procedures

Chapter 18 Comparing Two Means

Comparing two population means

Two-sample *t* procedures

Robustness again

Avoid the pooled two-sample *t* procedures[i]

Avoid inference about standard deviations[i]

Chapter 19 Inference about a Population Proportion

The sample proportion p^$\hat{p}$

Large-sample confidence intervals for a proportion

Accurate confidence intervals for a proportion

Choosing the sample size[i]

Hypothesis tests for a proportion

Chapter 20 Comparing Two Proportions

Two-sample problems: proportions

The sampling distribution of a difference between proportions

Large-sample confidence intervals for comparing proportions

Accurate confidence intervals for comparing proportions

Hypothesis tests for comparing proportions

Relative risk and odds ratio[i]

Discussion: Assessing and understanding health risks

Chapter 21 The Chi-Square Test for Goodness of Fit

Hypotheses for goodness of fit

Expected counts and chi-square statistic

The chi-square test for goodness of fit

Interpreting significant chi-square results

Conditions for the chi-square test

Chapter 22 The Chi-Square Test for Two-Way Tables

Two-way tables

Hypotheses for two-way tables of counts

Expected counts and chi-square statistic

The chi-square test

Conditions for the chi-square test

The chi-square test and the two-sample *z* test[i]

Chapter 23 Inference for Regression

The regression parameters

Testing the hypothesis of no linear relationship

Testing lack of correlation[i]

Confidence intervals for the regression slope

Inference about prediction

Checking the conditions for inference

Chapter 24 One-Way Analysis of Variance: Comparing Several Means

Comparing several means

The *F* statistic

The analysis of variance *F* test

Conditions for ANOVA

Details of ANOVA calculations[i]

Chapter 25 Statistical Inference: Part III Review

Essential concepts from Parts I and II

Part III summary

Comprehensive review exercises: inference selection

Comprehensive review exercises: analysis and conclusion

Large data set exercises

Online data sources

EESEE case studies

Notes and Data Sources

Tables

TABLE A: Random Digits

TABLE B: Standard Normal Probabilities

TABLE C: *t* Distribution Critical Values

TABLE D: Chi-square Distribution Critical Values

TABLE E: Critical Values of the Correlation *r*

TABLE F: *F* Distribution Critical Values

Answers to Selected Exercises

Some Studies Recurring Across Chapters

Index

Symbols

Part IV Optional Companion Chapters (available on the *PSLS 4e* Companion Website)

Chapter 26 More about Analysis of Variance: Follow-up Tests and Two-Way ANOVA

Beyond one-way ANOVA

Follow-up analysis: Tukey pairwise multiple comparisons

Follow-up analysis: contrasts[i]

Two-way ANOVA: conditions, main effects, and interaction

Inference for two-way ANOVA

Some details of two-way ANOVA[i]

Chapter 27 Nonparametric Tests

Comparing two samples: the Wilcoxon rank sum test

Matched pairs: the Wilcoxon signed rank test

Comparing several samples: the Kruskal-Wallis test

Chapter 28 Multiple and Logistic Regression

Parallel regression lines

Estimating parameters

Using technology

Conditions for inference

Inference for multiple regression

Interaction

A case study for multiple regression

Logistic regression

Inference for logistic regression

TO THE INSTRUCTOR: **About this Book**

THE PRACTICE OF STATISTICS IN THE LIFE SCIENCES

The Practice of Statistics in the Life Sciences, Fourth Edition (*PSLS 4e*) is an introduction to statistics for college and university students interested in the quantitative analysis of life science problems. Statistics has become an integral part of the life sciences, where its practice faces a specific set of challenges, such as observational studies with confounding variables and experiments with limited sample sizes. Consequently, students can greatly benefit from a teaching of statistics that is explicitly applied to their major. Fortunately, tailoring an introductory statistics course to a specific audience does not require fundamentally altering the statistical content if interesting and relevant examples are used to illustrate the statistical concepts.[1]

All examples and exercises in *PSLS* describe actual studies drawn from diverse areas of biology, such as physiology, brain and behavior, epidemiology, health and medicine, nutrition, ecology, and microbiology. Each new edition brings in recently published studies, a number of which cover topics of acute relevance. *PSLS 4e*, for instance, includes new problems on the Zika virus, the long-term effects of concussions, and patterns in global climate change. Instructors can choose to either cover a wide range of topics or select examples and exercises related to a particular field.

PSLS focuses on the applications of statistics rather than the mathematical foundation. This book is adapted from David Moore's best-selling introductory statistics textbook *The Basic Practice of Statistics* (*BPS*). *BPS* was the pioneer in presenting a modern approach to statistics in a genuinely elementary text. Like *BPS*, *PSLS* emphasizes balanced content, working with real data, and statistical ideas. It does not require any specific mathematical skills beyond being able to read and use simple equations and can be used in conjunction with almost any level of technology for computations and graphs.

GAISE GUIDING PRINCIPLES

The American Statistical Association (ASA) endorses a set of recommendations for introductory statistics courses at the college level described in the Guidelines for Assessment and Instruction in Statistics Education (GAISE). Initially published in 2005,[2] the GAISE College Report was revised in 2016 to account for the "many changes in the world of statistics education and statistical practice since 2005."[3] Previous editions of *PSLS* followed the initial GAISE recommendations closely, and the fourth edition has been revised specifically with the 2016 GAISE update in mind. Here are the six GAISE recommendations for the teaching of introductory statistics and how they are implemented in *PSLS 4e*.

1. **Teach statistical thinking.** An introductory statistics course should help students develop the skills required to think critically about statistical issues and to approach statistical problems like a statistician. Statistical thinking should be taught as an investigative process that recognizes the need for data, the importance of data production, the omnipresence of variability, the quantification and explanation of variability, and the relevance of multivariate exploration.

   To this end, *PSLS* begins with data analysis (Chapters 1 to 5) and data production (Chapters 6 and 7) in Part I, then moves on to probability (Chapters 9 to 12) and the concepts of statistical inference (Chapters 13 to 15) in Part II, before eventually covering a variety of inference procedures (Chapters 17 to 25 and companion Chapters 26 to 28) in Parts III and IV. This sequence matches the investigative process of statistical problems and facilitates students' learning experience by starting with topics that are more familiar to them.

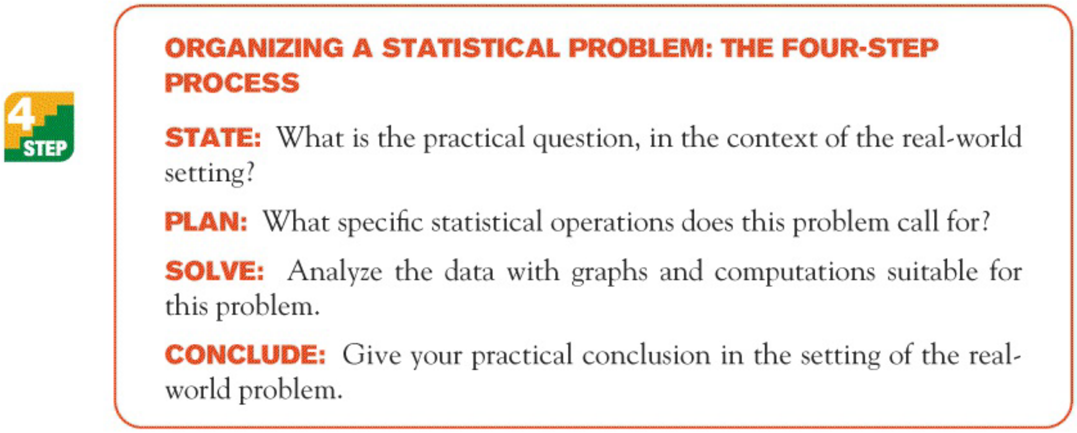
   It is easy for students to become lost in the details of the procedures covered in separate chapters. In the end, however, students should come to realize that these various parts are all interconnected and necessary to solving statistical problems. To help them along, each chapter contains a summary section titled "**This Chapter in Context**" that specifically highlights how the concepts from the chapter relate to concepts introduced in earlier chapters and how they will figure into following chapters. For a wider perspective, a **Discussion** box in Chapter 15 describes the general statistical approach used to solve scientific problems in the context of the Nobel Prize–winning discovery of the bacterial origin of most peptic ulcers. This detailed historical and scientific account helps students see how the concepts they learn throughout the book come together to form a coherent science of data.

Pedagogically, many of the examples and exercises throughout *PSLS* are presented in the context of a “**four-step process**” (State, Plan, Solve, Conclude), depicted in Figure 1. This process models the whole statistical approach for students: Statistical problems originate in a real-world setting (“State”) and require conclusions in the language of that setting (“Conclude”). Translating the problem into the formal language of statistics (“Plan”) is a key to success. The graphs and computations needed (“Solve”) are essential but are not the whole story. The four-step process (identified by a margin icon) appears throughout the text so as to foster the students’ ability to investigate statistical problems independently.

In this fourth edition, a greater emphasis has been placed, whenever appropriate, on the multivariate nature of statistics and the reliance on models. This is not to say that the topics covered in *PSLS* or the level at which they are covered has changed. *PSLS* remains an introductory statistics textbook meant to be easily accessible to students. Rather, the text explicitly points out multivariate patterns in context, to better familiarize students with this important aspect of statistical thinking.

2. **Focus on conceptual understanding.** A first course in statistics introduces many skills and it is important to ensure that students gain a deep understanding of the concepts behind them rather than a procedural knowledge of how to use a series of formulas like recipes. In practice (even if not always in the course), calculations and graphs are automated. Moreover, anyone who makes serious use of statistics will need some specific procedures not taught in his or her college statistics course. Thus, students should leave their first course with an understanding of the basic ideas of statistics, including the need for data, the importance of data production, the omnipresence of variability, the quantification and explanation of variability, and the

limitations of statistical procedures.



ORGANIZING A STATISTICAL PROBLEM: THE FOUR-STEP PROCESS

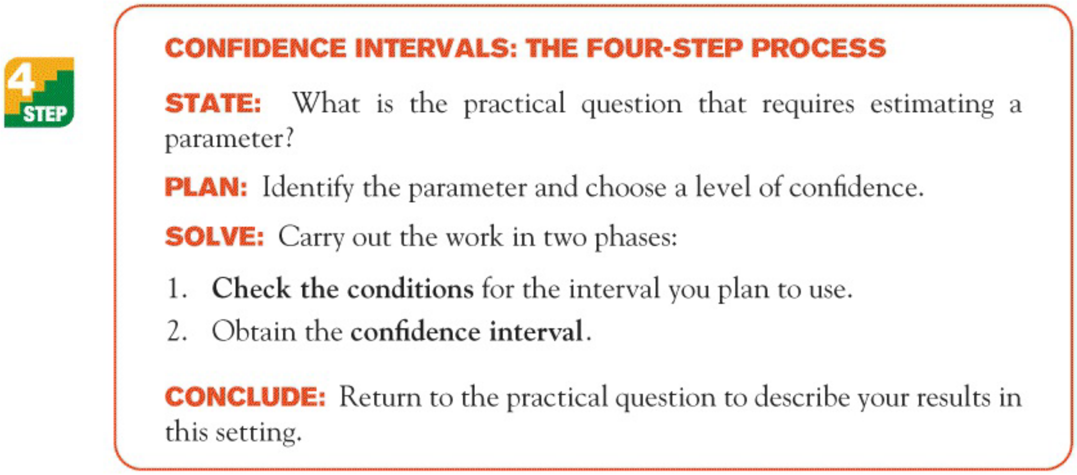
STATE: What is the practical question, in the context of the real-world setting?

PLAN: What specific statistical operations does this problem call for?

SOLVE: Analyze the data with graphs and computations suitable for this problem.

CONCLUDE: Give your practical conclusion in the setting of the real-world problem.

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



CONFIDENCE INTERVALS: THE FOUR-STEP PROCESS

STATE: What is the practical question that requires estimating a parameter?

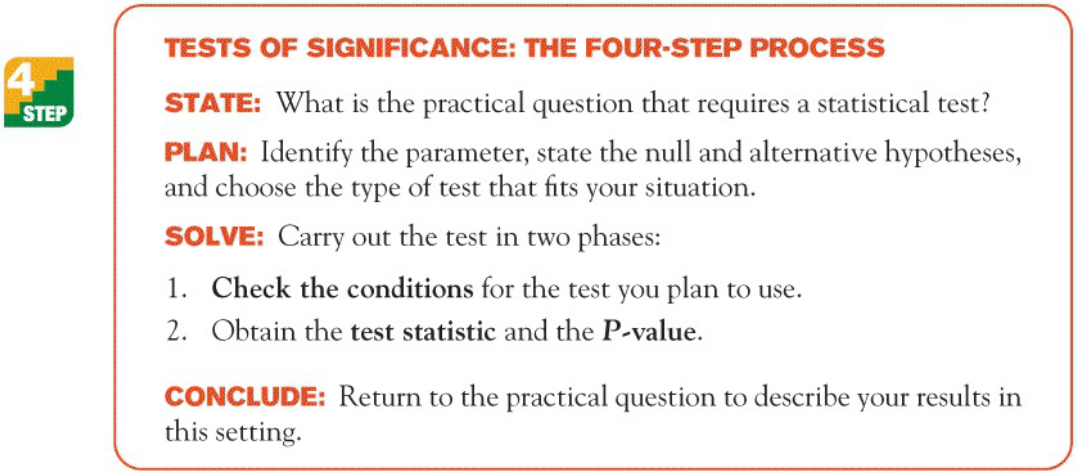
PLAN: Identify the parameter and choose a level of confidence.

SOLVE: Carry out the work in two phases:

1. **Check the conditions** for the interval you plan to use.
2. Obtain the **confidence interval**.

CONCLUDE: Return to the practical question to describe your results in this setting.

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



TESTS OF SIGNIFICANCE: THE FOUR-STEP PROCESS

STATE: What is the practical question that requires a statistical test?

PLAN: Identify the parameter, state the null and alternative hypotheses, and choose the type of test that fits your situation.

SOLVE: Carry out the test in two phases:

1. **Check the conditions** for the test you plan to use.
2. Obtain the **test statistic** and the ***P*-value**.

CONCLUDE: Return to the practical question to describe your results in this setting.

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1 Overview of the “four-step process” used in PSLS.

PSLS aims to make clear the larger patterns and big ideas of statistics in the context of learning specific skills and working with data. Many of the big ideas are summarized in graphical outlines in **Statistics in Summary** figures within the **Review Chapters**. The Review Chapters at the end of Parts I, II, and III offer a comprehensive summary of the important concepts and skills that students should have mastered. Review Chapters now also provide more comprehensive exercises that require applying concepts learned across a number of chapters.

Throughout the text, numerous cautionary statements are included to warn students about common confusions and misinterpretations. A handy “Caution” icon in the margin calls attention to these warnings. In addition, two Discussion boxes address how a meaningful interpretation must rely on a comprehensive analysis of the data available: The Discussion box in Chapter 10 addresses the interpretation of conditional probabilities in the context of diagnostic and screening tests, while the Discussion box in Chapter 20 helps students assess and interpret health risks beyond the *P*-value of a hypothesis test.

CAUTION

3. **Integrate real data with a context and purpose.** The study of statistics is supposed to help students work with data in their varied academic disciplines and later employment. This is particularly important for students in the life sciences, because they are asked to collect and analyze data in their laboratory courses and elective undergraduate research. *PSLS* prepares students by providing real (not merely realistic) data from many areas of the life sciences, with sources cited at the back of the book. Data are more than mere numbers—they are numbers with a context that should play a role in making sense of the numbers and in stating conclusions. Examples and exercises in *PSLS* give enough background to allow students to consider the meaning of their analysis in context.

   PSLS provides approximately 50 examples and exercises per chapter, ranging in complexity from studies with summary statistics or small data sets for in-class use to large data sets (with several variables and a fairly large number of subjects) more suitable for lab work or assignments. In addition, two Discussion boxes address in greater depth some important issues when dealing with real data: The Discussion box in Chapter 1 exposes

   LARGE DATA SET

students to some of the challenges of data entry and validation, while the Discussion box in Chapter 2 explains how to recognize different types of outliers and how to deal with them legitimately.

Real and relevant studies are also important to illustrate the usefulness of statistics and to generate enthusiasm in students.[4] A very small number of examples and exercises in *PSLS* are, for specific pedagogical purposes, theoretical or purely computational. All others describe actual life science studies drawn from scientific and news reports published in the years immediately before each new edition. The studies are selected for both their statistical relevance and general interest. This latest edition, for instance, addresses the U.S. opioid health crisis, the pathology of the Zika virus and possible treatment, Magellanic penguins, and the neurological basis for the infamous photo of a dress that looks black-and-blue to some people and white-and-gold to others.

4. **Foster active learning.** How students learn in the classroom is the domain of the instructor. *PSLS* offers a number of opportunities to help instructors foster active learning. After a summary of the chapter's key concepts, a set of **Check Your Skills** multiple-choice items with answers in the back of the book lets students assess their grasp of basic ideas and skills. These problems can also be used in a "clicker" classroom response system to enable class participation. Additional **clicker questions** are available to instructors on the book's Companion Website.

   PSLS also provides many examples and exercises based on small data sets or summary statistics that can be solved during class in a reasonable amount of time. The Review Chapters offer links to regularly updated **online data sources** (including some with built-in data visualization tools) to support classroom discussion. Instructors interested in creating a review session at the end of the term can build on a set of short exercises in Chapter 25 requiring students to select an inference procedure appropriate for the particular data, design, and purpose of various studies. For courses that include a computer lab component, the large data set exercises offer an opportunity for hands-on analysis with statistical software. There is no short answer given to students for these specific exercises, so that instructors can elect to assign them for a grade.

Visual representations of new concepts can also help students learn through experience. On the *PSLS* Companion Website, instructors and students will find a set of interactive **applets** created to our specifications. These applets are designed primarily to help students in learning statistics, rather than in doing statistics, and can be effective tools for classroom demonstrations of concepts such as the influence of outliers and sample size on various computations.

5. **Use technology to explore concepts and analyze data.** Technology is now an integral part of the practice of statistics and this reality should be reflected in its teaching. As an added bonus, automating graphs and calculations increases students' ability to complete problems that use real data (including those with larger, more complex data sets) and helps them focus on concepts and problem recognition rather than mechanics.

   PSLS does not require or recommend any particular technology. Instead, outputs from a variety of platforms and programs (applets, CrunchIt!®, Excel, JMP, Minitab, R/Rcmdr, SPSS[i], TI-83/84) are shown throughout the book, with similarities and notable differences pointed out, to help students familiarize themselves with **technology outputs**. The choice of technology and the extent of its use in the classroom and for assignments are often dictated by practical and financial considerations specific to each course and, therefore, are best left to the instructor. *PSLS 4e* now provides on its Companion Website technical instructions organized by chapter for CrunchIt!®, Excel, JMP, Minitab, R, SPSS, and the TI-83/84 in support of that choice. All **data** sets in *PSLS* are also available electronically in various formats to facilitate students' active learning.

   APPLET

   A quite different use of technology appears in the interactive applets available on the Companion Website. Some examples and exercises in the text use applets, and they are marked with a dedicated icon in the margin.

6. **Use assessments to improve and evaluate student learning.** *PSLS* provides a wide variety of exercises that can be used to create a mix of formative and summative assessments. Within chapters, exercises progress from straightforward applications (Apply Your Knowledge

short exercises and Check Your Skills multiple-choice items) to comprehensive end-of-chapter exercises. Complementing this set, exercises in the three Review Chapters enlarge the statistical context beyond that of the immediate lesson. (Many instructors will find that the Review Chapters appear at convenient points for pre-exam review.)

Short answers to odd-numbered exercises are provided to students (at the back of the book or via a click in the e-Book) for self-assessment. At the opposite end of the spectrum, exercises marked with the "Large Data Set" icon are never offered with short answers (regardless of placement) so that they may be used for graded assignments or in-class activities. These exercises rely on larger and slightly more complex data sets designed to provide a richer data experience to students. A few of these exercises are located at the end of individual chapters but, for the most part, they can now be found in Review Chapters to facilitate comprehensive analyses that are not limited by chapter boundaries.

LARGE DATA SET

WHAT'S NEW IN THE FOURTH EDITION?

This fourth edition of *PSLS* offers many new examples and exercises, as well as numerous opportunities to refocus the exposition to align with developments in statistics education. Here are, specifically, some of the major changes:

- **New exercises and examples.** As always, *PSLS* aims to provide an abundance and variety of current and interesting examples and exercises. The fourth edition has more than 400 new or updated exercises and examples, representing roughly 30% of the material in the third edition's main chapters. On the practical side, having many new exercises means that homework assignments can be varied over time to avoid "recycled paper" issues. To further help with this issue, Review Chapters in *PSLS 4e* include a set of links to regularly updated online data sources relevant to the life sciences that can be used to create easily renewed assignments.

  New exercises and examples also help instructors avoid the boredom of teaching the same old material and help students better appreciate the relevance of the statistical approach when it is applied to current issues. Some of the new problems with a medical or epidemiological focus include studies of the Zika virus (ranging from the association with microcephaly to a potential treatment), the U.S. opioid epidemic, the relationship between concussion and long-term brain damage, and approaches to smoking cessation. Behavioral studies include an ethically questionable Facebook experiment and and a study of the infamous photo of a dress that looks black-and-blue to some people and white-and-gold to others. *PSLS 4e* also presents studies of ant speeds and snow geese flight, adds two new large data sets on animal morphology (carpenter ants and monarch butterflies), describes plastic–eating bacteria that may help reduce plastic pollution, and covers several studies related to patterns of global climate change.

Beyond the studies themselves, the actual exposition of the concepts in examples and the type of questions asked in exercises have undergone some subtle yet important changes in *PSLS 4e*. The reason for this is pedagogical: Statistics is an evolving discipline and the field has seen substantial pedagogical updates in recent years. The 2016 GAISE update and the corresponding approach taken in *PSLS 4e* have already been addressed in depth. An additional development that occurred in 2016 was the release by the American Statistical Association (ASA) of an official statement on *P*-values. Following is a description of the statement and the associated changes in *PSLS 4e*.

- **Treatment of statistical significance.** Some misuses of statistical ideas have over time culminated in a problem of reproducibility and replicability of scientific conclusions and an eroding trust in statistics. Because many of these issues centered on *P*-values, the ASA published an official statement on *P*-values.[5] This 2016 ASA statement is brief, written to be accessible beyond the statistical community, and makes a great reading assignment for students in an introductory statistics course (Exercise 16.17, page 410, is specifically designed for this purpose). The statement clarifies what *P*-values do and do not imply, and points to their appropriate versus inappropriate uses.

  Of particular concern is the fairly common reliance (among students, scientists, and publishers) on “P<0.05$P < 0.05$” as a sort of “bright-line” threshold bestowing scientific worth. In a self-reinforcing loop, this practice became ever more widespread as statistics courses taught $\alpha=0.05$ precisely because it was so widespread. In an attempt to break the cycle, *PSLS 4e* explicitly de-emphasizes the omnipresence of the significance level α (0.05 or other), expecting instead that students evaluate the strength of the evidence against the null hypothesis. Students should understand that, when the conditions for inference are met, and given similar study settings, a *P*-value of 0.0006 provides stronger evidence against H_0 than a *P*-value of 0.047, and that neither is a judgment of practical importance.

Beyond issues of significance level, the true meaning of *P*-values is often misunderstood. In this edition, the actual definition of a *P*-value is explicitly stated in every inference chapter to remind students of what can and cannot be concluded from a *P*-value and that *P*-values come with assumptions that must be checked. These caveats are not mere side notes; rather, they are fundamental to the use of statistical inference.

- **“Using Technology.”** Previous editions of *PSLS* covered the use of technology for computations in a separate section titled “Using Technology,” typically located toward the end of the chapter. *PSLS 4e* employs a different approach. Technology is now part of all examples for which it is relevant, and it is fully integrated with a discussion of study design, choice of methodology, and conclusion in context. Computations are still outlined and explained explicitly for every example, but they no longer occupy the center stage. Furthermore, the use of tables of critical values is now demonstrated in optional sections within their relevant chapters, to serve courses that rely on them as an alternative to technology in the classroom.[6]

  The new approach models for students a more realistic practice of statistics. It also lets students focus on all the truly challenging aspects of statistics (such as stating an appropriate and comprehensive conclusion in context) that no amount of technology can accomplish for them.

- **Parts and Review Chapters.** The flow of topics in *PSLS 4e* remains unchanged, but the organization of topics into parts reflects a more thematic content-grouping that is also likely to better fit the format of a statistics course with regular projects or exams. Specifically, the chapters on data acquisition (formerly Chapters 7 and 8) have been brought into Part I, renamed “Collecting and Exploring Data.” Part II is now titled “From Chance to Inference” and is more fundamentally focused on issues related to probability.

The new grouping reflects the changing nature of our student population, which is more likely to have already encountered many of the concepts of descriptive statistics covered in Chapters 1 through 5. To this end, Chapters 1 and 2 in particular have been updated to de-emphasize hand-computations of summary statistics and to present a greater variety of graphical displays that students are likely to encounter in news reports and scientific publications.

The new chapter grouping also creates more thematically homogeneous Review Chapters, allowing for comprehensive problems that draw from several chapters at once. Similarly, most Large Data Set Exercises (including several brand-new data sets) have been placed in the Review Chapters to facilitate comprehensive analyses that are not limited by chapter boundaries. Lastly, for instructors who like to assign projects, links to regularly updated online data sources round up the collection.

- **Discussion exercises.** Instructors have told us that, while they found the various Discussion boxes throughout *PSLS* very interesting, they also found them challenging to officially “assign.” The fourth edition now contains exercises following each Discussion box that specifically ask students to consider which aspects of the discussion relate to a particular problem.
- **Technology appendices.** Over the years, numerous instructors have indicated that short technology instructions would be helpful for students learning to use a particular technology for their assignments. We are excited to announce that the Companion Website for *PSLS 4e* now provides technical instructions, organized by chapter, for CrunchIt!®, Excel, JMP, Minitab, R, SPSS, and the TI-83/84. For instructors using our online homework platforms, video instructions remain available as well.

WHY DID YOU DO THAT?

There is no single best way to organize the presentation of statistics to beginners. That said, our choices reflect thinking about both content and pedagogy. Here are comments on several “frequently asked questions” about the order and selection of material in *PSLS 4e.*

Why not begin with data production? It is certainly reasonable to do so—the natural flow of a planned study is from design to data analysis to inference. In their future employment, though, most students will use statistics mainly in settings other than planned research studies. Students also come into an introductory statistics course having some experience with the basics of descriptive statistics, so it makes sense pedagogically to start with this topic.

Why not delay correlation and regression until late in the course? *PSLS 4e* begins by offering experience working with data and gives a conceptual structure for this nonmathematical but essential part of statistics. Students profit from more experience with data and from seeing the conceptual structure worked out in relationships among variables as well as describing single-variable data. Correlation and least-squares regression are very important descriptive tools, and they are often used in settings where there is no population-sample distinction, such as studies based on state records or average species data. Perhaps most important, the *PSLS* approach asks students to think about what kind of relationship lies behind the data (confounding, lurking variables, association doesn’t imply causation, and so on), without overwhelming them with the demands of formal inference methods. The same argument can be made for introducing two-way tables first in the context of exploratory data analysis. Of course, instructors sometimes need to adapt their syllabus to suit nonpedagogical issues such as course length and sequence. The writing in *PSLS* is modular enough that instructors can easily pick a set of chapters and an order that works for them.

What about probability? Chapters 9, 11, and 13 present—in a simple format—the ideas of probability and sampling distributions that are needed to understand inference. These chapters go from the idea of probability as long-term regularity, to the concrete use of basic probability models, to the idea of the sampling distribution of a statistic. The central limit theorem appears in the context of discussing the sampling distribution of a sample mean. The *optional* Chapters 10 and 12 cover the more advanced topics of independence and conditional probability, and the binomial and Poisson distributions.

Depending on your student audience, you may want to omit one or both of these optional chapters. Experienced teachers recognize that students find probability difficult, and research has shown that this is true even for professionals. If a course is intended for medical school or premed students, for instance, the concept of conditional probability is very relevant because it is a key part of diagnosis that both doctors and patients have difficulty interpreting.[7] However, attempting to present a substantial introduction to probability in a data-oriented statistics course for students who are not mathematically trained is a difficult challenge. Instructors should keep in mind that formal probability does not help students master the ideas of inference as much as we teachers often imagine, and it depletes reserves of mental energy that might better be applied to essential statistical ideas.

Why use the *z* procedures for a population mean to introduce the reasoning of inference? This is a pedagogical issue, rather than a question of statistics in practice. At some point in the golden future, we may start with resampling methods, as permutation tests make the reasoning of tests clearer than any traditional approach. For now, the main choices are *z* for a mean and *z* for a proportion.

The *z* procedures for means are pedagogically more accessible to students. We can say up front that we will explore the reasoning of inference in an overly simple setting. Remember, exactly Normal populations and true simple random samples are as unrealistic as known σ, especially in the life sciences. All the issues of practice—robustness against lack of Normality and application when the data aren't a simple random sample, as well as the need to estimate σ—are put off until, with the reasoning in hand, we discuss the practically useful *t* procedures. This separation of initial reasoning from messier practice works well.

In contrast, starting with inference for *p* introduces many side issues: no exact Normal sampling distribution, but a Normal approximation to a discrete distribution; use of p^$\hat{p}$ in both the numerator and the denominator of the test statistic to estimate both the parameter *p* and p^$\hat{p}$'s own standard deviation; loss of the direct link between test and confidence interval. In addition, we now know that the traditional z confidence interval for p is often grossly inaccurate, as explained in the following section. Lastly, major polling organizations such as Gallup and Pew now report substantially different margins of error (accounting for data weighing), making it increasingly challenging to show students real-life applications of the z method for a proportion.

Furthering changes initiated in previous editions, *PSLS 4e* covers the z procedures for a population mean at a mostly conceptual level, so as to introduce the reasoning of inference. Instructors may choose to pair these conceptual chapters with Chapter 17 on the t procedures for a population mean, thereby combining theory and practice.

Why does the presentation of inference for proportions go beyond the traditional methods? Recent computational and theoretical work has demonstrated convincingly that the standard confidence intervals for proportions can be trusted only for very large sample sizes. It is hard to abandon old friends, but the graphs in section 2 of the paper by Brown, Cai, and DasGupta in the May 2001 issue of *Statistical Science* are both distressing and persuasive.[8] The standard intervals often have a true confidence level much less than what was requested, and requiring larger samples encounters a maze of "lucky" and "unlucky" sample sizes until very large samples are reached. Fortunately, there is a simple cure: Just add two successes and two failures to your data. (Therefore, no additional software tool is required for this procedure.) We present these "plus four intervals" in Chapters 19 and 20, along with guidelines for use.

Why haven't the companion Chapters 26 to 28 been updated at this time? A new edition takes more than a year to edit and ready for print (in addition to finding new studies and writing new material). Because the companion chapters aren't printed, we have chosen to focus first on the core text chapters and the synchronous production of our homework system. The companion chapters will be updated between editions, and made available on the Companion Website.

Why didn't you cover Topic X? Introductory texts ought not to be encyclopedic. Including each reader's favorite special topic results in a text that is formidable in size and intimidating to students. The topics covered in *PSLS* were chosen because they are the most commonly used in the life sciences and they are suitable vehicles for learning broader statistical ideas. Three chapters available on the Companion Website cover more advanced inference procedures. Students who have completed the core of *PSLS* will have little difficulty moving on to more elaborate methods.

Acknowledgments

We are grateful to colleagues from two-year and four-year colleges and universities who commented on successive drafts of the manuscript. Special thanks are due to Marjorie Bond of Monmouth College, who read the manuscript and checked its accuracy. We also wish to thank those who reviewed the fourth edition:

Reza O. Abbasian,
Texas Lutheran University
James Robert Adcock,
University of Waterloo
Amei Amei,
University of Nevada, Las Vegas
Alexander Aue,
University of California, Davis
Steve Bacinski,
Davenport University
Jennifer Beineke,
Western New England University
Margaret Bryan,
University of Missouri, Columbia
Grace Chen,
Idaho State University
Tyler Cook,
University of Central Oklahoma
William M. Cook,
St. Cloud State University
Dina Dawoud,
University of Waterloo
Lelys Bravo de Guenni,
University of California, Santa Cruz
Christiana Drake,

University of California, Davis

Donna Dueker,
California State University, Dominguez Hills

Diane Fisher,
University of Louisiana at Lafayette

Leszek Gawarecki,
Kettering University

Jose Guardiola,
Texas A&M University?Corpus Christi

Abeer Hasan,
Humboldt State University

Leslie Hendrix,
University of South Carolina

Heather J. Hoffman,
George Washington University

Dawn Holmes,
University of California, Santa Barbara

Yi Huang,
University of Maryland, Baltimore County

Michelle Jeitler,
Marietta College

Zhenyu Jia,
University of Akron

Ravi Khattree,
Oakland University

Jong S. Kim,
Portland State University

Yoon G. Kim,
Humboldt State University

Elizabeth Young Kolodziej,
Texas A&M University

Chris Lacke,
Rowan University

Maureen Lahiff,
University of California, Berkeley
Suhwon Lee,
University of Missouri, Columbia
Lia Liu,
University of Illinois at Chicago
Tatiana Mihaylova,
Bellevue College
Christopher Mosbo,
State University of New York at Oswego
Linda Myers,
Harrisburg Area Community College
Leonard Onyiah,
St. Cloud State University
Adriana Perez,
University of Texas Health Science Center at Houston
Amber Powell,
State University of New York at Fredonia
Tiantian Qin,
Purdue University
Rachel Rader,
Ohio Northern University
Laurence D. Robinson,
Ohio Northern University
Julie Roy,
University of Louisiana at Lafayette
Laura M. Schultz,
Rowan University
Ye Shen,
University of Georgia
J. McLean Sloughter,
Seattle University
Dawit G. Tadesse,
University of Cincinnati

Pamini Thangarajah,
Mount Royal University
Kyle T. Turner,
University of Georgia
Jennifer Waugh,
University of Western Ontario
Daniel Weiner,
Boston University
Qiang Wu,
East Carolina University

We are also particularly grateful to Karen Carson, Tom DeMarco, Katrina Mangold, Andy Newton, Lisa Kinne, Paul Rohloff, Robin Fadool, Naomi Kornhauser, Marco Baldi, and the other editorial and design professionals who have helped make this textbook a reality.

Finally, our thanks go to the many students whose compliments and complaints have helped improve our teaching over the years.

Brigitte Baldi and David S. Moore

Media and Supplements

INTRODUCING SAPLINGPLUS FOR PSLS 4E!

The new fourth edition of *PSLS* is now in SaplingPlus, an extraordinary new online resource providing the richest, most completely integrated text/media learning experience yet. SaplingPlus combines Macmillan's StatsTools, powerful multimedia resources, and text-specific machine-gradable exercises for *PSLS* with an integrated e-Book and the robust exercise library of Sapling Learning.

Created and supported by educators, **Macmillan Learning's Instructional Online Homework** drives student success and saves educators time. More than 1000 exercises contain full solutions and general feedback, and more than 800 exercises also contain hints, answer-specific feedback, and solutions, ensuring that students find the help they need, wherever they need it. With SaplingPlus, every problem counts.

ASSETS INTEGRATED INTO SAPLINGPLUS INCLUDE THE FOLLOWING:

Interactive e-Book provides powerful study tools for students, video and multimedia content, and easy customization for instructors. Students can search, highlight, and bookmark specific information, making it easier to study and access key content.

LearningCurve provides students and instructors with powerful adaptive quizzing, a game-like format, direct links to the e-Book, and instant feedback. The quizzing system features questions tailored specifically to the text and adapts to students' responses, providing material at different difficulty levels and topics based on student performance.

StatTutor Tutorials offer multimedia tutorials that explore important concepts and procedures in a presentation that combines video, audio, and interactive features. The newly revised format includes built-in assignable assessments and a bright new interface.

Statistical Applets give students hands-on opportunities to familiarize themselves with important statistical concepts and procedures in an interactive setting that allows them to manipulate variables and see the results graphically. Icons in the textbook indicate when an applet is available for the material being covered.

StatBoard Videos are brief whiteboard videos that illustrate difficult topics through additional examples, written and explained by an expert group of statistics educators.

Statistical Video Series consists of StatClips, StatClips Examples, and Statistically Speaking "Snapshots." View animated lecture videos, whiteboard lessons, and documentary-style footage that illustrate key statistical concepts and help students visualize statistics in real-world scenarios.

NEW! Technology Appendices for each chapter give detailed instructions on how to perform key statistical procedures in various statistical software packages including TI-83/84 calculators, Minitab, Excel, JMP, SPSS, R, Rcmdr, and CrunchIt!®. Also available on the *PSLS* Companion Website.

Video Technology Manuals, available for TI-83/84 calculators, Minitab, Excel, JMP, SPSS, R, Rcmdr, and CrunchIt!®, provide brief instructions for using specific statistical software.

CrunchIt!® is Macmillan Learning's own web-based statistical software that allows users to perform all the statistical operations and graphing needed for an introductory statistics course and more. It saves users time by automatically loading data from *PSLS*, and it provides the flexibility to edit and import additional data.

JMP Student Edition (developed by SAS) is easy to learn and contains all the capabilities required for introductory statistics. JMP is the leading commercial data analysis software of choice for scientists, engineers, and analysts at companies throughout the world (for Windows and Mac).

Data files are available in JMP, ASCII, Excel, TI, Minitab, SPSS, R, and CSV formats.

EESEE Case Studies (*Electronic Encyclopedia of Statistical Examples and Exercises*) developed by The Ohio State University Statistics Department, teach students to apply their statistical skills by exploring actual case studies using real data.

Student Solutions Manual provides solutions to the odd-numbered exercises in the text.

Instructor's Guide with Solutions includes teaching suggestions, chapter comments, and detailed solutions to all exercises. Available to instructors only.

Test Bank offers hundreds of multiple-choice questions. Available to instructors only.

Lecture Slides offer a customizable, detailed lecture presentation of statistical concepts covered in each chapter of *PSLS*. Available to instructors only.

ADDITIONAL RESOURCES AVAILABLE WITH PSLS:

Companion Website www.macmillanlearning.com/psls4e

This open-access website includes:

- Statistical applets
- Data files
- Companion chapters covering further topics in ANOVA, nonparametric tests, and multiple and logistic regression
- Selected *EESEE* Case Studies that are referenced in the text
- Technology Appendices for each chapter

Instructor access to the Companion Website requires user registration as an instructor and features all the open-access student web materials, plus:

- Instructor's Guide with Solutions
- Downloadable Test Bank specific to each chapter in *PSLS*

- Lecture Slides specific to each chapter in *PSLS*
- iClicker Questions specific to each chapter in *PSLS*
- Image Slides containing all textbook figures and tables

Special Software Package: Student versions of JMP are available for packaging with the text. JMP is available inside SaplingPlus at no additional cost. Contact your Macmillan Learning representative for information or visit www.macmillanlearning.com.

CrunchIt!®: With a course adoption, instructors may provide CrunchIt!® access at no cost to students. Please contact your Macmillan Learning representative to receive a course access code for CrunchIt!®

i-clicker is a two-way radio-frequency classroom response solution developed by educators for educators. Each step of i-clicker’s development has been informed by teaching and learning. To learn more about packaging i-clicker with this textbook, please contact your local sales representative or visit www.iclicker.com.

TO THE STUDENT: **Statistical Thinking**

Statistics is about data. Data are numbers, but they are not “just numbers.” **Data are numbers with a context.** The number 7.5, for example, carries no information by itself. But studies of sleep patterns[1] indicate that a child in middle school who sleeps 7.5 hours on most nights is likely not getting enough sleep, yet that same amount of sleep is appropriate for most adults. The context engages our background knowledge and allows us to make judgments; it makes the number informative.

Statistics is the science of data. To gain insight from data, we analyze the data with various graphs and computational approaches. In essence, graphs and calculations are guided by ways of thinking that amount to educated common sense. Let’s begin our study of statistics with an informal look at some principles of statistical thinking.

DATA BEAT ANECDOTES

An anecdote is a striking story that sticks in our minds exactly because it is striking. Anecdotes humanize an issue, but they can be misleading.

Does living near power lines cause leukemia in children? The National Cancer Institute spent 5 years and $5 million gathering data on this question. The researchers compared 638 children who had leukemia with 620 who did not. They went into the homes and measured the magnetic fields in the children’s bedrooms, in other rooms, and at the front door. They recorded facts about power lines near the family home and also near the mother’s residence when she was pregnant. Result: No connection between leukemia and exposure to magnetic fields of the kind produced by power lines. The editorial that accompanied the study report in the *New England Journal of Medicine* thundered, “It is time to stop wasting our research resources” on the question.[2]

Now compare the effectiveness of a television news report of a 5-year, $5 million investigation against a televised interview with an articulate mother whose child has leukemia and who happens to live near a power line. In the public mind, the anecdote wins every time. This is why, for instance, many parents worldwide refuse to vaccinate their children against potentially devastating childhood diseases, despite the countless studies showing *no association between vaccines and autism.*[3] A statistically literate person knows better. **Data are more reliable than anecdotes, because they systematically describe an overall picture rather than focus on a few incidents.**

WHERE THE DATA COME FROM IS IMPORTANT

The advice columnist Ann Landers once asked her readers, “If you had it to do over again, would you have children?” A few weeks later, her column was headlined “70% OF PARENTS SAY KIDS NOT WORTH IT.” Indeed, 70% of the nearly 10,000 parents who wrote in said they would not have children if they could make the choice again. Do you believe that 70% of *all* parents regret having children?

You shouldn’t. The people who took the trouble to write Ann Landers are not representative of all parents. Their letters showed that many of them were angry at their children. All we know from these data is that there are some unhappy parents out there. A statistically designed poll, unlike Ann Landers’s appeal, targets specific people chosen in a way that gives all parents the same chance to be asked. Such a poll showed that 91% of parents *would* have children again. Where data come from matters a lot. If you are careless about how you get your data, you may announce 70% “no” when the truth is about 90% “yes.”

Here’s another question: *Should episiotomy be a routine part of childbirth?* Episiotomy is surgical cutting of the skin and muscles between the vagina and the rectum that is sometimes performed during childbirth. Until recently, it was one of the most common surgical procedures performed in women in the United States. The idea was that a sharp incision ought to be better for the mother than the possibility of a natural tearing.

However, recent clinical trials and epidemiological studies have shown no benefit of episiotomy unless the baby’s health requires accelerated delivery or a large natural tearing seems likely. In fact, these studies indicate that episiotomy is associated with longer healing times and increased rates of complications, including infection, extensive tearing, pain, and incontinence.[4]

To get convincing evidence on the benefits and risks of episiotomy, we need unbiased data. Proper epidemiological studies and clinical trials (controlled, randomized medical experiments) rely on randomness of patient or treatment selection to avoid bias. These careful studies of the risks and benefits of routine episiotomy can be trusted because they had sound data collection designs. As a result, in 2006, the American Congress of Obstetricians and Gynecologists recommended against routine episiotomy.

The most important information about any statistical study is how the data were produced. Only surveys that are statistically designed can be trusted. Only randomized experiments can give convincing evidence that an alleged cause really does account for an observed effect.

BEWARE THE LURKING VARIABLE

Energy drinks contain high levels of caffeine, which we know can temporarily boost alertness but can also cause sleep problems later on. A survey of U.S. service members in a combat environment in Afghanistan showed that those consuming energy drinks three or more times a day slept less, had more sleep disruptions, and were more likely to fall asleep on duty than those consuming fewer or no energy drinks.[5] What shall we make of this finding? Among service members in combat deployment, does the frequent consumption of energy drinks impair sleep or do individuals with impaired sleep consume more energy drinks to stay alert on the job? The answer may well be both, in a self-reinforcing loop. It is also possible that other factors, such as high stress, both interfere with sleep and create a desire for heightened alertness from energy drinks. A statistician knows that a causal link cannot be established here, because the data from this survey are simply observations.

Should women take hormones such as estrogen after menopause, when natural production of these hormones ends? In 1992, several major medical organizations said “yes.” In particular, women who took hormones seemed to reduce their risk of a heart attack by 35% to 50%. But in 2002, the National Institutes of Health declared these findings wrong.

Use of hormones after menopause immediately plummeted. Both recommendations were based on extensive studies. What happened?

The evidence in favor of hormone replacement came from a number of *observational studies* that compared women who were taking hormones with others who were not. But women who choose to take hormones are very different from women who do not: They are richer and better educated and see doctors more often. These women do many things to maintain their health. It isn't surprising that they have fewer heart attacks.

We can't conclude that hormone replacement reduces heart attacks, just because we see a relationship in the data. Education and affluence are *lurking variables,* background factors that help explain the relationship between hormone replacement and good health.

Almost all relationships between two variables are influenced by other variables lurking in the background. To understand the relationship between two variables, you must often look at other variables. Careful statistical studies try to think of and measure possible lurking variables so as to correct for their influence. As the hormone saga illustrates, this is not always done. News reports often ignore possible lurking variables that might ruin a good headline. The habit of asking "What might lie behind this relationship?" is part of thinking statistically.

Another way to address the effect of lurking variables is by designing careful *experiments.* Several experiments randomly assigned volunteer women either to hormone replacement or to dummy pills that looked and tasted the same as the hormone pills. By 2002, these experiments agreed that hormone replacement does *not* reduce the risk of heart attacks, at least for older women. Faced with this better evidence, medical authorities changed their recommendations.[6]

ALWAYS LOOK AT THE DATA

Yogi Berra said it best: "You can observe a lot by just watching." That's a motto for learning from data. **A few carefully chosen graphs are often more instructive than great piles of numbers.**

Let's look at some data. Figure 1 displays a histogram of the body lengths of 56 perch (*Perca fluviatilis*) caught in a Finnish lake.[7] Each bar in the graph represents how many perch had a body length between two values on the horizontal axis. For example, the tallest bar indicates that 17 perch had a body length between 20 and 25 centimeters (cm).

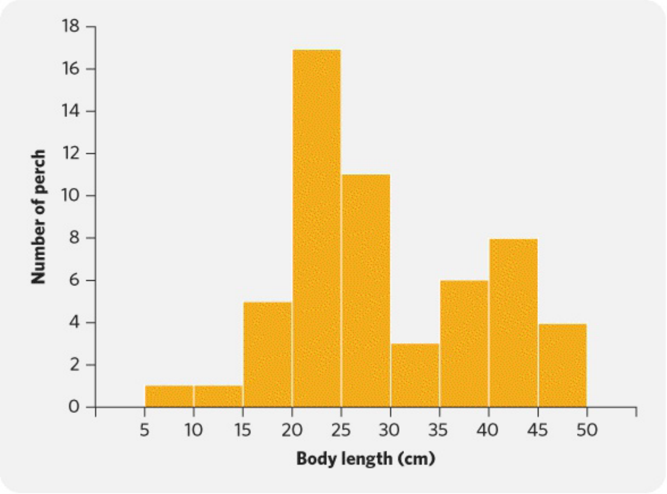


Figure 1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1 Body lengths of 56 perch from a Finnish lake. Always look at the data: Two separate clusters are clearly visible.

We see great variability in perch body lengths, from about 5 to 50 cm, but we also notice two clear clusters: One group of smaller fish and one group of larger fish. This suggests that the data include two different groups of perch, possibly males and females or perch of different ages. In fact, observations in the wild indicate that perch can become much larger if they survive to a very old age (up to 22 years).

Any attempt to summarize the data in Figure 1 without a close inspection first would miss the two distinct clusters. Failure to examine any data carefully can lead to misleading or absurd results. As humorist Des McHale put it, “The average human has one breast and one testicle.”

VARIATION IS EVERYWHERE

Many students in biology lab courses are surprised to obtain somewhat different results when they repeat an experiment. Yet, variability is ubiquitous.

Figure 2 plots the ventilation rates of seven goldfish placed in tanks of varying temperature.[8] Each goldfish is represented by one color. While there is a clear overall pattern in this figure, there is also a lot of variability, with ventilation rates ranging overall from 6 to 119 opercular movements per minute. Some of that variability can be attributed to lack of precision in measurements, some to fish physiology, and some to slight differences in circumstances at the time of each measurement (such as movements around the fish tanks that might stress the fish). Before the experiment, the fish were kept in tanks whose water temperature was set to 22 degrees Celsius. At that temperature, the fish ventilation rates range from about 60 to 100 opercular movements per minute. Yet when the same seven fish are transferred to containers with warmer water, their ventilation rates increase overall. And when the fish are transferred to containers with water at lower temperatures, the ventilation rates decrease quite dramatically. These data show that ventilation rate in goldfish varies both from fish to fish and as a function of water temperature.

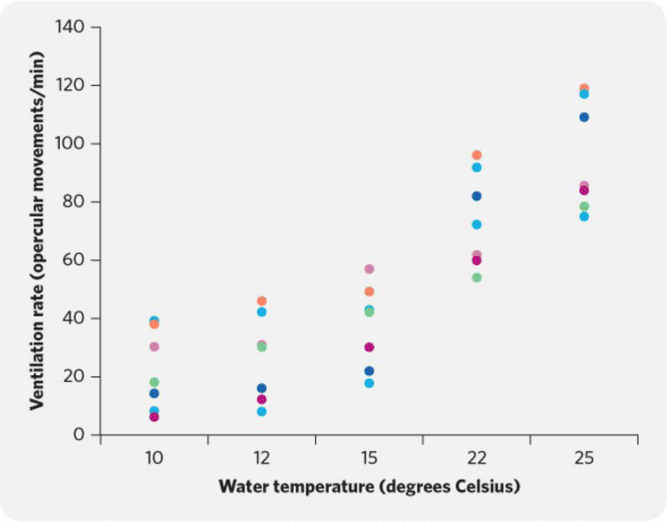


Figure 2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2 Variation is everywhere: The ventilation rates of seven goldfish placed in tanks with varying water temperature.

Students are not the only people with a tendency to underestimate variability in real data. Here is Arthur Nielsen, head of the country’s largest market research firm, describing his experience:

> *Too many business people assign equal validity to all numbers printed on paper. They accept numbers as representing Truth and find it difficult to*

> *work with the concept of probability. They do not see a number as a kind of shorthand for a range that describes our actual knowledge of the underlying condition.*[9]

Variation is everywhere. Individuals vary; repeated measurements on the same individual vary; almost everything varies over time. One reason we need to know some statistics is that statistics helps us deal with variation and uncover underlying patterns.

CONCLUSIONS ARE NOT CERTAIN

Most women who reach middle age have regular mammograms to detect breast cancer. *Do mammograms reduce the risk of dying of breast cancer?* Doctors rely on clinical trials that compare different ways of screening for breast cancer. The conclusion of the U.S. Preventive Services Task Force after examining all the data available is that biennial mammograms reduce the risk of death in women aged 50 to 69 years by about 16.5% compared with no screening.[10]

Because variation is everywhere, conclusions are uncertain. Statistics gives us a language for talking about uncertainty that is used and understood by statistically literate people everywhere. On average, women who have mammograms every two years between the ages of 50 and 69 are less likely to die of breast cancer. But because variation is everywhere, the results are different for different women. Some women who have biennial mammograms die of breast cancer, and some who never have mammograms live long, breast-cancer-free lives. Statistical conclusions are "on average" statements only. In addition, the conclusion that mammograms reduce the risk of death from breast cancer comes from experiments involving some, and not all, women aged 50 to 69 years. In fact, the various studies found a reduction in breast cancer deaths versus no screening ranging from 15% to 23%. So can we be certain that mammograms reduce risk on average? No. We can be very confident, but we can't be absolutely certain. This is an important distinction that is often not understood. We will learn how to quantify our level of confidence in statistical conclusions.

WHAT LIES AHEAD IN THIS BOOK

The purpose of *The Practice of Statistics in the Life Sciences* (*PSLS*), Fourth Edition, is to give you a working knowledge of the ideas and tools of practical statistics. We will divide practical statistics into three main areas:

1. **Data.** Data analysis concerns methods and strategies for exploring, organizing, and describing data using graphs and numerical summaries. Only organized data can illuminate reality. Only thoughtful exploration of data can defeat the lurking variable. Chapters 1 to 5 in Part I of *PSLS* cover data analysis for different types of data and different types of patterns between variables. Different methods for collecting data can be used to give clear answers to specific questions. Where the data come from really is important. Basic concepts about how to select samples and design experiments are the most influential ideas in statistics. These concepts are the subject of Chapters 6 and 7 within Part I.
2. **Probability.** Probability is the language that describes variation and uncertainty. Because we are concerned with practice, rather than theory, we need only a limited knowledge of probability. Chapters 9, 11, and 13 in Part II of *PSLS* introduce the most important probability rules and models, complemented by Chapters 10 and 12 for those students whose major requires a more in-depth coverage. Chapters 14 and 15 in Part II apply probability concepts to data analysis, taking into account that variation is everywhere and that conclusions are uncertain. These chapters present the reasoning of statistical inference in depth but at a conceptual level, so that what you learn there applies to all following chapters covering various statistical inference procedures.
3. **Statistical inference.** While the concepts of statistical inference have wide applications, the actual computations involved vary by data types and study designs. Chapters 17 to 24 in Part III of *PSLS* present inference procedures used in simple, common settings. The review Chapter 25 helps you recognize when to use each of these procedures. The optional Chapters 26 to 28, found on the Companion Website, cover more advanced or specialized kinds of inference that may be commonly used in some fields of study.

Statistics, however, is an integrated science of data, and you will see that concepts from one area often relate directly to concepts from other areas. To help you see this, each chapter contains a summary section titled "**This Chapter in Context**." These sections highlight how the concepts from the current chapter relate to concepts introduced in earlier chapters and how they will figure in following chapters.

Because data are numbers with a context, doing statistics means more than manipulating numbers. You must **state** a problem in its real-world context, **plan** your specific statistical work in detail, **solve** the problem by making the necessary graphs and calculations, and **conclude** in context by explaining what your findings say about the real-world setting. We'll make regular use of this **four-step process** (easily recognizable by its margin icon) to encourage good habits that go beyond graphs and calculations to ask, "What do the data tell me?"

Statistics does involve lots of calculating and graphing. This text offers detailed explanations so that their specific use and limitations can be better understood. In practice, however, all computations and graphs in modern statistics are automated with technology. The proper use of a particular technology is a technical issue unrelated to the big ideas of statistics, so *PSLS 4e* teaches how to *read* various **technology** outputs but not which command to type or which button to press. The Companion Website, however, provides technical instructions, organized by chapter, for CrunchIt!®, Excel, JMP, Minitab, R, SPSS, and the TI-83/84 graphing calculator.

Because graphing and calculating are automated in statistical practice, the most important benefits you can gain from the study of statistics are an understanding of the big ideas and the beginnings of good judgment in working with data. *PSLS* aims to explain the most important ideas of statistics, not just to teach methods. Some examples of big ideas that you will encounter in this text are "always plot your data," "randomized comparative experiments," and "statistical significance." Some particularly important ideas (such as how to treat outliers and the scientific approach) are given more extensive treatment with real-life examples in **Discussion** boxes placed throughout the book.

You learn statistics by doing statistical problems. As you read, you will see several levels of exercises, arranged to help you learn. Short **Apply Your Knowledge** problem sets appear after each major idea. These straightforward exercises will help you solidify the main points as you read. Make sure you can do these exercises before going on. The end-of-chapter exercises begin with multiple-choice **Check Your Skills** exercises (with all answers in the back of the print book or within a click in the e-Book): Use them to check your grasp of the basics. The regular **Chapter Exercises** help you combine all the ideas of a chapter. Throughout the book, exercises marked with the Large Data Set icon provide an opportunity to apply your statistical skills to the kind of large and complex data sets often encountered in the life sciences. In addition, the three Part **Review chapters** look back over major blocks of learning, offering review exercises that provide a more realistic practice of statistics that is not artificially limited to the confines of a chapter.

LARGE DATA SET

The Part Review chapters (and the individual optional chapters on the Companion Website) also include point-by-point lists of specific things you should be able to do. They can be very useful for self-guided study and review sessions. Finally, because ease with statistics is earned through practice, consider doing more problems than just those officially assigned in your class. Every odd-numbered exercise in the book (except for exercises marked “Large Data Set”) has a short answer available (at the back of printed books or by clicking a link in the e-Books). You can use this feature to check your work and evaluate your progress.

The key to learning is persistence. The main ideas of statistics, like the main ideas of any important subject, took a long time to discover and take some time to master. The rewards from mastering these concepts will be well worth the effort. In the words of Jessica Utts, 2016 president of the American Statistical Association, “People are realizing that statistical literacy is an important component of being a well-educated person and is relevant to daily life.”

About the Authors

Brigitte Baldi is a graduate of France's École Normale Supérieure in Paris. In her academic studies, she combined a love of math and quantitative analysis with wide interests in the life sciences. She majored in math and biology and obtained a master's degree in molecular biology and biochemistry and a master's degree in cognitive sciences. She earned her PhD in neuroscience from the Université Paris VI, studying multisensory integration in the brain. As a postdoctoral fellow at the California Institute of Technology, she used computer simulations to study patterns of brain reorganization after lesion. She then worked as a management consultant advising corporations, before returning to academia to teach statistics.

Dr. Baldi is currently a lecturer in the Department of Statistics at the University of California, Irvine. She is actively involved in statistical education. She was a local and later national adviser in the development of the statistics telecourse *Statistically Speaking*, a new take on David Moore's earlier telecourse *Against All Odds.* She developed UCI's first online statistics courses and is always interested in ways to integrate new technologies in the classroom to enhance participation and active learning. She is a member of the American Statistical Association, currently serving as webmaster for the Section on Statistical Education.

David S. Moore is Shanti S. Gupta Distinguished Professor of Statistics, Emeritus, at Purdue University and was 1998 president of the American Statistical Association. He received his AB from Princeton and his PhD from Cornell, both in mathematics. He has written many research papers in statistical theory and served on the editorial boards of several major journals. Professor Moore is an elected fellow of the American Statistical Association and of the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. He has served as program director for statistics and probability at the National Science Foundation.

In recent years, Professor Moore has devoted his attention to the teaching of statistics. He was the content developer for the

Annenberg/Corporation for Public Broadcasting college-level telecourse *Against All Odds: Inside Statistics* and for the series of video modules *Statistics: Decisions through Data*, intended to aid the teaching of statistics in schools. He is the author of influential articles on statistics education and of several leading texts. Professor Moore has served as president of the International Association for Statistical Education and has received the Mathematical Association of America’s national award for distinguished college or university teaching of mathematics.

PART I

Reinhard Dirscherl/Getty Images

The first step in understanding data is to hear what the data say, to "let the statistics speak for themselves." Numbers tell a story in context, but only after we organize, display, and summarize them. That's *exploratory data analysis*:

- The purpose is unrestricted exploration of the data to search for interesting patterns and unusual observations.
- Conclusions are informal, based on what we see in the data, and they apply only to the individuals and circumstances for which we have data in hand.

The first five chapters in Part I present the ideas and tools of modern statistical data analysis. This requires first looking at variables one at a time and then examining their relationships. In Chapters 1 and 2 you will study *variables and their distributions*. Chapters 3, 4, and 5 concern *relationships among variables*.

Informative data analysis requires good data production because how the data are collected impacts what we can learn from them. Chapters 6 and 7 present the statistical ideas of *data production* to answer specific questions.

Chapter 8 reviews this part of the text and provides more comprehensive exercises that cross chapter boundaries.

Collecting and Exploring Data

EXPLORING DATA: VARIABLES AND DISTRIBUTIONS

Chapter 1 Picturing Distributions with Graphs
Chapter 2 Describing Quantitative Distributions with Numbers

EXPLORING DATA: RELATIONSHIPS

Chapter 3 Scatterplots and Correlation
Chapter 4 Regression
Chapter 5 Two-Way Tables

PRODUCING DATA

Chapter 6 Samples and Observational Studies
Chapter 7 Designing Experiments

Chapter 8 Collecting and Exploring Data: Part I Review

CHAPTER 1 Picturing Distributions with Graphs

mark@rocketclips.com/Depositphotos

IN THIS CHAPTER WE COVER...

- Individuals and variables
- Identifying categorical and quantitative variables
- Categorical variables: pie charts and bar graphs
- Quantitative variables: histograms
- Interpreting histograms
- Quantitative variables: dotplots
- Time plots
- *Discussion: (Mis)adventures in data entry*

Statistics is the science of data. Data collection has increased tremendously in the new millennium, and this trend shows no signs of slowing down. In 2003, the Human Genome Project uncovered, after 13 years of international cooperation and billions of dollars, the complete sequence of the 3 billion DNA bases of the human genome. A decade later, individuals can have their own genome sequenced within weeks for just thousands of dollars. In the United States, the Census Bureau collects extensive data on the nation from numerous surveys, such as the yearly National Health Interview Survey, which gathers health and socioeconomic information on each member of 40,000 households. Gallup, a private polling firm, surveys roughly 1000 U.S. adults per day for its Gallup-Healthways Well-Being Index covering a whole range of physical and mental health conditions. Smartphones have become so ubiquitous that they are used to monitor ongoing epidemics worldwide. Satellites provide detailed daily records of climatic conditions at the planetary level. Millions of individuals now routinely use apps and wearables to collect and store a wealth of health data, which are then carefully examined by private individuals and large corporations alike. “Big data” is an exploding phenomenon that has brought about a renewed interest in data exploration. While some of the techniques used with extremely large and complex data sets are very specific to their unique needs, the fundamental principles are essentially the same as those we will study in this and the following few chapters.

INDIVIDUALS AND VARIABLES

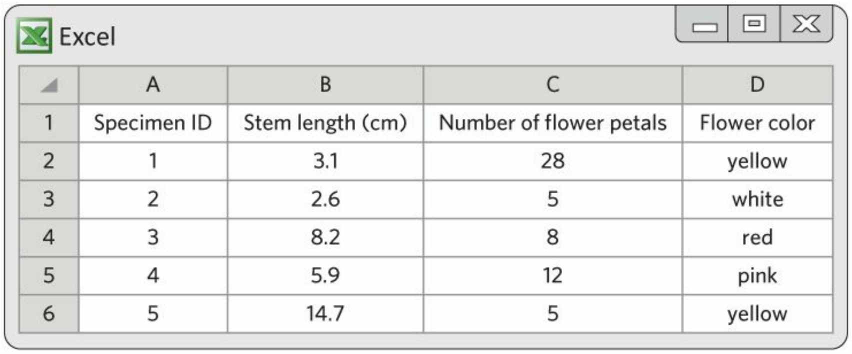
Any set of data contains information about some group of *individuals*. The information is organized in *variables*. The techniques of descriptive statistics covered in Part I of this book apply equally to data sets obtained from a given **population** (the entire group of individuals about which we want information) and to data sets collected from only those individuals in a smaller **sample**. The distinction between a population and a sample is an important one for inference, and we will address it in more detail in Chapter 6.

INDIVIDUALS AND VARIABLES

Individuals are the objects (or units) described by a set of data. Individuals may be people, but they may also be animals, plants, or things.

A **variable** is any one characteristic of an individual. A variable can take different values for different individuals.

A botanist's plant database, for example, includes data about various aspects of the plants examined. The plants are the individuals described by the database. For each individual (each plant), the data contain the values of variables such as stem length, number of flower petals, and flower color. An example of what such a database might look like is shown in Figure 1.1, with each row dedicated to one individual and each column to a variable. **Spreadsheet** programs with rows and columns readily available are commonly used to store data and do simple calculations.



	A	B	C	D
1	Specimen ID	Stem length (cm)	Number of flower petals	Flower color
2	1	3.1	28	yellow
3	2	2.6	5	white
4	3	8.2	8	red
5	4	5.9	12	pink
6	5	14.7	5	yellow

Figure 1.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.1 Example of a botanical database displayed in a spreadsheet. Each column includes data about a different variable. Each row represents data for one plant specimen.

Note that *sometimes the individuals represented in a data set are actually groups*, such as a set of countries, states, or counties, or even a set of ant colonies. For instance, every year, the U.S. Centers for Disease Control and Prevention (CDC) reports the percent of obese individuals in each state. In this data set, the 50 states are the individuals studied and the variable recorded for each individual is the percent of the population in that state who are obese.

CAUTION

IDENTIFYING CATEGORICAL AND QUANTITATIVE VARIABLES

Some variables, like a plant’s flower color, simply place individuals into categories. Others, like stem length and number of flower petals, take numerical values with which we can do arithmetic. It makes sense to give an average stem length for plants of a given environment, but it does not make sense to give an “average” flower color. We can summarize our findings on plant color by obtaining the counts of yellow, pink, and white flowers in the database, or by obtaining the proportion that each color type represents. *It is important, though, not to confuse these numerical summaries (average, count, or proportion) with the variables themselves.*

CATEGORICAL AND QUANTITATIVE VARIABLES

A **categorical variable** places an individual into one of several groups or categories.

A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense. The values of a quantitative variable are usually recorded in a **unit of measurement** such as seconds or kilograms.

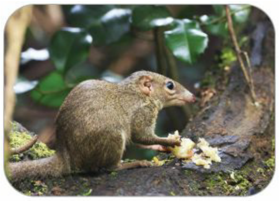
Further distinctions are sometimes made in variables beyond simply categorical or quantitative. Some quantitative variables, like stem length, are **continuous** variables that can take any real numerical value over an interval. **Discrete** variables, in contrast, are quantitative variables that can take only a limited, finite number of values, like the number of petals in a flower.[1] Categorical variables can also be broken down into nominal and ordinal variables. **Nominal** variables are purely qualitative and unordered, like flower color, whereas **ordinal** data can be ranked, like star ratings or the Likert scales commonly used in psychology (for example, “Do you strongly disagree, disagree, agree, or strongly agree with the following statement?”). Although ordinal data can be ranked, they are not true quantitative variables, because the intervals between consecutive ranks are often not identical.

Be sure to carefully identify the individuals in the study first, so that you can determine what was recorded for each individual. For example, when the CDC reports the percent of obese individuals in each state, the individuals studied are the 50 states—not people. While people are either obese or not obese, each *state* provides a meaningful numerical value (the percent of its population who are obese). Therefore, the variable here is "percent of the population who are obese," and it is a quantitative variable.

EXAMPLE 1.1 Paw preference in tree shrews

Tree shrews, *Tupaia belangeri*, are small omnivorous mammals that are phylogenetically related to primates. A research team examined paw preference during grasping tasks among 36 tree shrews born and raised in captivity. Here are some of the data they reported:[2]

Subject	Sex	Age	PI	Bias
Abel	m	6	−1.00	L
Anna	f	5	−0.95	L
Aragorn	m	6	−1.00	L
Barbossa	m	1	−0.25	A
Bea	f	4	1.00	R
Beatrice	f	1	0.56	R
Berta	f	1	0.13	A
...				



Oscar Dominguez/age fotostock/
SuperStock

In addition to the animal's name (an identifier), four variables are displayed. The first one, sex, is a categorical variable because each animal can be only male (m) or female (f). *Sex*

CAUTION

> *would still be a categorical variable if it had been recorded with a “0” for male and a “1” for female, because these would be labels rather than true numerical values.* The second variable, age, is a quantitative variable measured in years. Age values displayed in the table are rounded to whole numbers of years, but animals age continuously: thus, age is continuous and not discrete. The third variable, PI, stands for “pawedness index,” a unitless continuous variable taking values between −1 and 1, depending on the relative use of the right and left paws. The last variable, bias, is categorical and represents the animal’s general paw preference: right paw (R), left paw (L), or ambidextrous (A).

Few scientific publications display their raw data the way the study in Example 1.1 does. Instead, information about a study’s individuals and variables must often be retrieved from the abstract, the methods section, or a figure legend, typically presented in summary form. When you explore data from someone else’s work—or when planning your own statistical study—ask the following questions:

1. **Who** or what are the **individuals** studied? **How many** individuals appear in the data set?
2. **What** do we record for each individual? How many **variables** do the data contain? What are the **exact definitions** of these variables? Are they categorical or quantitative? In what **units of measurement,** if any, is each variable recorded? Lengths, for example, might be recorded in inches, in yards, or in meters.
3. **Why? What purpose** do the data have? Do we hope to answer some specific questions? Do we want to draw conclusions about individuals other than the ones we actually have data for? Are the variables suitable for the intended purpose?

STATS IN YOUR WORLD

GESUNDHEIT!

Researchers who study treatments for the common cold must find ways to assess their effectiveness. Effectiveness may be defined as either preventing a cold or reducing its symptoms. Researchers can record simply whether or not each person exposed to the common-cold virus developed cold symptoms. They can also count the number of cold symptoms a person experienced, or count the number of tissues a person used in a day, or record the duration of symptoms. Smart studies often look at a problem from more than just one angle.

EXAMPLE 1.2 Tumor-specific antibodies for cancer treatment

Could cancer cells be selectively targeted by using antibodies recognizing a tumor-specific protein marker? Researchers grafted human cancerous cells onto healthy adult mice. Some of the mice were then injected with tumor-specific antibodies (treatment group, 10 mice), while the other mice were not (control group, 10 mice). A figure in the publication displays the following findings: “the number of mice exhibiting lymph node metastases in each group” and “the number of secondary lymph nodes detected in each mouse.”[3]

First we ask who or what are the individuals studied. Here the individuals are the mice injected with human cancerous cells. There were 20 mice in total, divided into two groups of 10 each. Then we ask what was recorded for each individual. The figure describes two findings. While both findings make reference to a “number of …,” the first number represents the summary finding for each group, whereas the second number describes individual mice. Therefore, the first variable is the presence or absence of lymph node metastases, a categorical variable. The second variable is the number of secondary lymph nodes, a quantitative discrete variable. Lastly we ask what purpose the data serve. The first variable answers the question of whether the treatment can prevent the formation of metastases (yes or no), while the second variable answers the question of how effective the treatment is (how many nodes are found).

As Example 1.2 illustrates, it is important not to jump to conclusions. While words like “number,” “count,” “proportion,” or “percent” may be used to summarize categorical data, seeing these words does not necessarily imply that the variable recorded is categorical.

The publication referenced in Example 1.2 presented a lot more findings. This is not particularly unusual. *Some scientific publications combine the findings from a series of related but distinct studies.* In such cases, you need to identify each study before you can determine the specific individuals and variables addressed in the study.

CAUTION

APPLY YOUR KNOWLEDGE

1.1 Cereal content. Here is a small part of an EESEE data set (available on the companion website), “Nutrition and Breakfast Cereals,” that describes the nutritional content per serving of 77 brands of breakfast cereal:[4]

Brand name	Manufacturer	Cold/Hot	Calories (Cal)	Sugar (grams)	Fiber (grams)
All Bran	K	C	70	5	9
All Bran with Extra Fiber	K	C	50	0	14
Almond Delight	R	C	110	8	1
Apple Cinnamon Cheerios	G	C	110	10	2
Apple Jacks	K	C	110	14	1
⋮					

a. What are the individuals in this data set?

b. For each individual, what variables are given? Which of these variables are categorical and which are quantitative?

1.2 Concussion and sleep. A 2016 National Public Radio (NPR) report described a study comparing patients who had their first ever traumatic brain injury 18 months earlier with similar healthy individuals with no prior brain trauma.[5] The news report states that, 18 months after their traumatic brain injury, the patients “were still getting, on average, an hour more sleep each night than similar healthy people were getting. And despite the extra sleep, 67 percent showed signs of excessive daytime sleepiness. Only 19 percent of healthy people had that problem.”

a. Identify the individuals in the study.
b. Identify the variables recorded and whether they are categorical or quantitative.

mark@rocketclips.com/Depositphotos

CATEGORICAL VARIABLES:

Pie charts and bar graphs

Statistical tools and ideas help us examine data so we can describe their main features. This examination is called **exploratory data analysis.** Like an explorer crossing unknown lands, we want first to simply describe what we see. Here are two principles that help us organize our exploration of a set of data.

EXPLORING DATA

1. Begin by examining each variable by itself. Then move on to study the relationships among the variables.
2. Begin with a graph or graphs. Then add numerical summaries of specific aspects of the data.

We will also follow these principles in organizing our learning. Chapters 1 and 2 present methods for describing a single variable. We study relationships among several variables in Chapters 3 to 5. In each case, we begin with graphical displays, then add numerical summaries for more complete description.

The proper choice of graph depends on the nature of the variable. To examine a single variable, we usually want to display its *distribution.*

DISTRIBUTION OF A VARIABLE

The **distribution** of a variable tells us what values it takes and how often it takes these values.

The values of a categorical variable are labels for the categories. The **distribution of a categorical variable** lists the categories and gives either the count or the percent of individuals that fall in each category.

EXAMPLE 1.3 Infectious diseases

The California Health and Human Services offers an open data portal (at https://chhs.data.ca.gov) providing state data on a variety of topics. You can, for instance, access all cases of infectious diseases reported in California. The database shows that, in 2014, California had a total of 262,780 reported infectious cases. Here is how these cases break down by type of infectious disease, after lumping the rarer diseases into an "other" category:

Infectious disease	Number of cases	Percent of all cases
Chlamydia	174,557	66
Gonorrhea	44,974	17
Pertussis	11,219	4
Campylobacteriosis	7,919	3
Early syphilis	7,191	3
Salmonellosis	5,361	2
Other	11,559	4
Total	262,780	100

The data table provides both the number of reported cases for each type of infectious disease and the percent that each disease represents in the set. Counts are also sometimes referred to as **frequencies,** and percents as **relative frequencies.** This table lists only the most common infectious diseases and lumps the remaining cases in the "other" category. Note that in 2014 California experienced an unusually large epidemic of pertussis, also known as whooping cough. Pertussis was the third most common infectious disease reported that year.

It's a good idea to check data for consistency. In this data set, the number of cases should add to 262,780, the total number of infectious disease cases reported in California in the year 2014. They do. The percents should add to 100% or, because each percent in the table is rounded to the nearest integer, very nearly 100%. Here the rounded percents add to 99 rather than 100. **Roundoff errors** don't point to mistakes in our work, just to the effect of rounding off results.

Columns of numbers take time to read and interpret. The **pie chart** in Figure 1.2 shows the distribution of infectious cases more vividly. For example, the “gonorrhea” slice makes up 17% of the pie because 17% of all reported cases of infectious diseases in California in 2014 were gonorrhea cases. Pie charts are awkward to create by hand, but software will quickly do the job for you. *A pie chart can represent only one variable in one group at a time and must include all the categories that make up a whole.* Use a pie chart when you want to emphasize each category’s relation to the whole. Here the cases all fall into either one of the six most common infectious diseases categories or the “other” category that lumps together all other infectious diseases. The graph shows more clearly than the table of raw numbers the overwhelming predominance of chlamydia and gonorrhea, two sexually transmitted diseases.

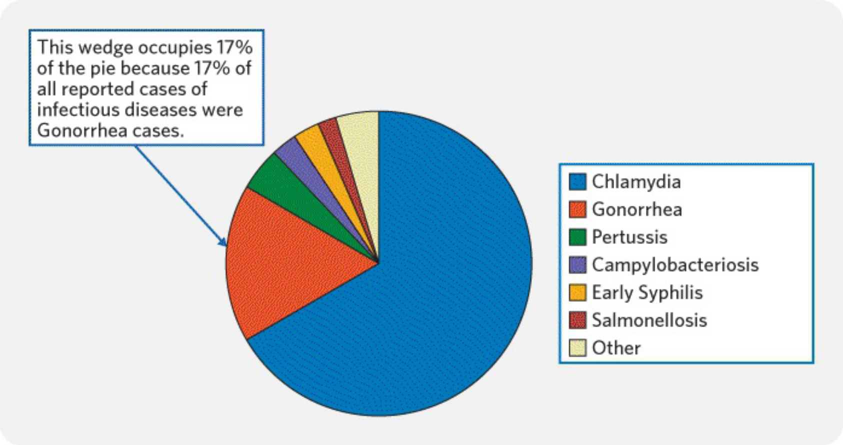
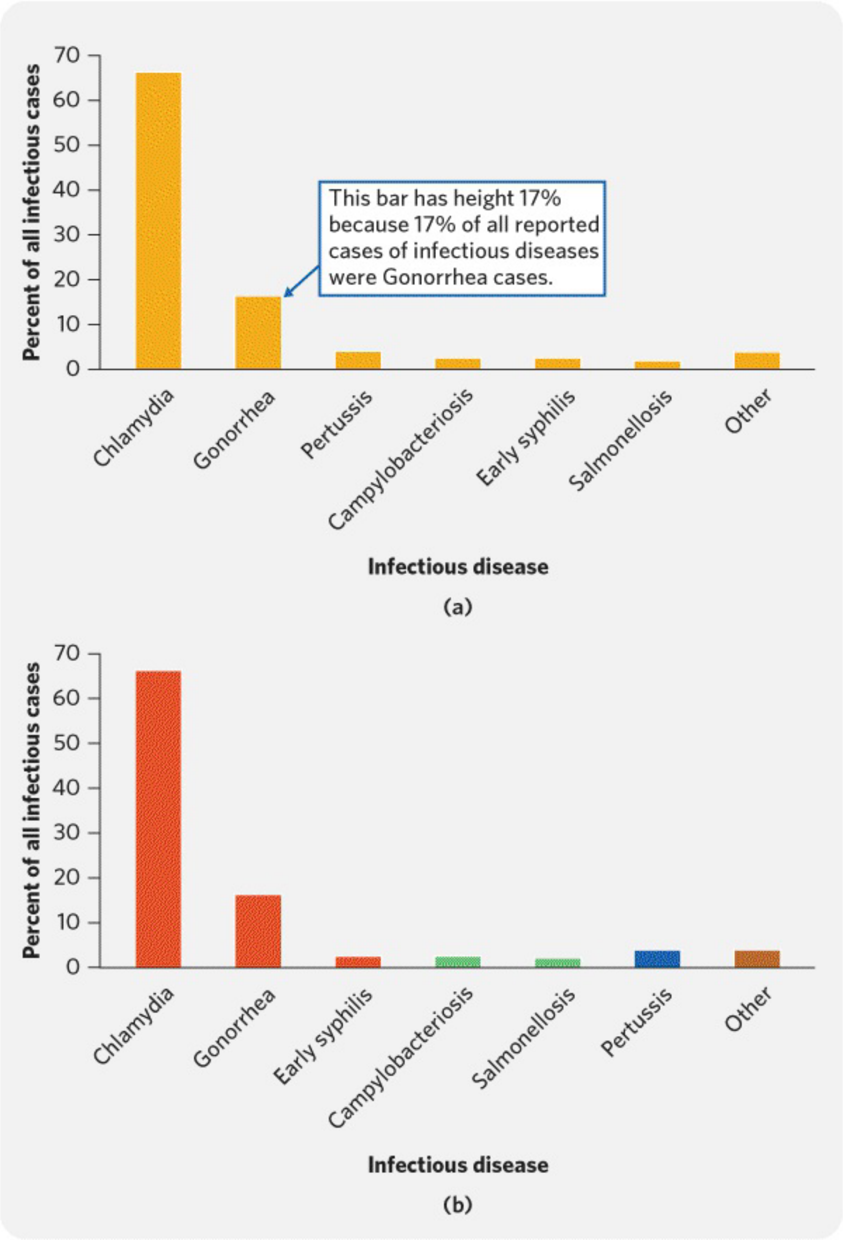


Figure 1.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.2 You can use either a pie chart or a bar graph to display the distribution of a categorical variable. Here is a pie chart of the reported cases of infectious diseases in California for 2014.

We could also make a **bar graph** that represents each infectious disease through the height of a bar. Bar graphs are particularly adept at pointing out the order and the relative importance of the different categories. Figure 1.3 shows two possible bar graphs of the data in Example 1.3. The bar graph shown in Figure 1.3(a) is sorted by decreasing order of magnitude: The tallest bar appears first, followed by the second-tallest bar, and so on. It makes very clear that chlamydia and gonorrhea were by far the most common cases of infectious diseases reported in California in 2014. Because the data here are categorical, the bars in the graph may be arranged in any order we choose. Although the result would not be particularly interesting, the bars could be sorted alphabetically. Alternatively, the bars could be grouped by type of disease, as in Figure 1.3(b). We can more easily see now that three of the six most common infectious diseases reported in California that year were sexually transmitted diseases (chlamydia, gonorrhea, syphilis) and that two were foodborne diseases (campylobacteriosis and salmonellosis).[i]



This bar has height 17% because 17% of all reported cases of infectious diseases were Gonorrhea cases.

Figure 1. 3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.3 The data on the reported cases of infectious diseases in California for 2014 are displayed in a bar graph format. (a) The bars are sorted according to their height (by order of importance). (b) The bars in the graph are grouped by disease type.

Pie charts must include all the categories that make up a whole (one variable in one group), but bar graphs are more flexible: They can be used to compare groups and do not necessarily display all possible outcomes of a variable. Beware that *this flexibility implies that bar graphs often are more challenging to interpret*

CAUTION

correctly.

EXAMPLE 1.4 Who is more likely to have suicidal thoughts?

The National Survey on Drug Use and Health reports the percent of adults in the United States who had serious suicidal thoughts in 2014, for each of four age groups:[6]

Age group (years)	Percent who had serious suicidal thoughts
18–25	7.5
26–44	4.2
45–64	3.5
65+	1.6

It’s clear that young adults between 18 and 25 years of age are the most susceptible to serious suicidal thoughts.

Figure 1.4(a) is a bar graph of the data in Example 1.4. We can see at a glance that, among adults, the proportion who had serious suicidal thoughts declines with age. We can’t make a pie chart to display the data in Example 1.4, because each percent in the table refers to a different age group, not to parts of a single whole.

The **stacked bar graph** in Figure 1.4(b) makes this point more clearly: In each age group, individuals who had serious suicidal thoughts and individuals who did not together make up 100% of that age group.

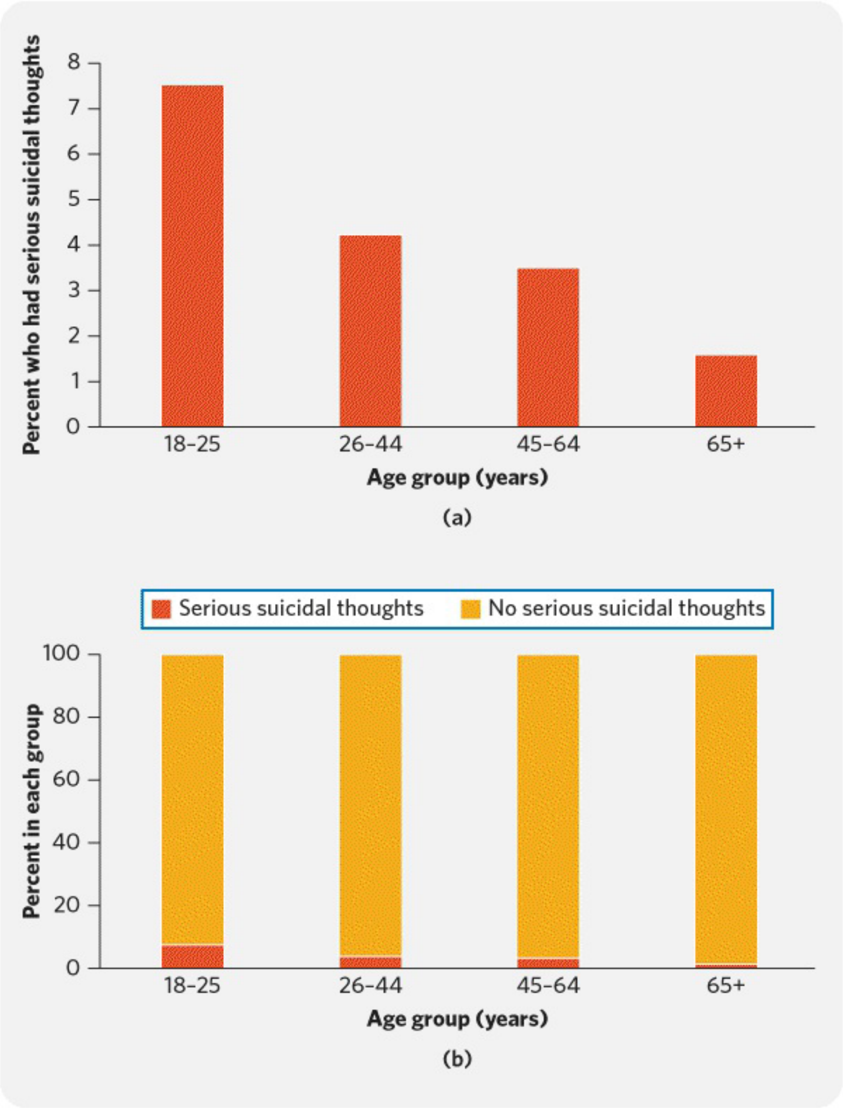


Figure 1.4

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.4 Bar graph showing the percent of American adults in each of four age groups who had serious suicidal thoughts in 2014. (a) The graph displays only those individuals in each group who had serious suicidal thoughts. (b) The graph displays both individuals who did and did not have serious suicidal thoughts in each age group, so that the stacked bars make up 100% of each age group.

It's important to understand this point to interpret the data correctly. Notice that the 18 to 25 age group represents a much smaller share of the American adult population than the other age groups listed, and 7.5% of this smaller age group could constitute a relatively small number of individuals who had serious suicidal thoughts. So we can conclude from the data that young adults aged 18 to 25 have the highest rate of serious suicidal thoughts, but not necessarily that they represent a large fraction of all individuals who have serious suicidal thoughts.

The ability to juxtapose data for different groups is what makes bar graphs so informative. The data in Example 1.4 were published in the Behavioral Health Barometer, United States, 2015, using the graph in Figure 1.5. This graph shows the percent of individuals who had serious suicidal thoughts, broken down side by side by age group and also by race/ethnicity. The graph makes it clear that age group is an influential factor but that, by comparison, race/ethnicity has very little effect on the disposition to have serious suicidal thoughts.

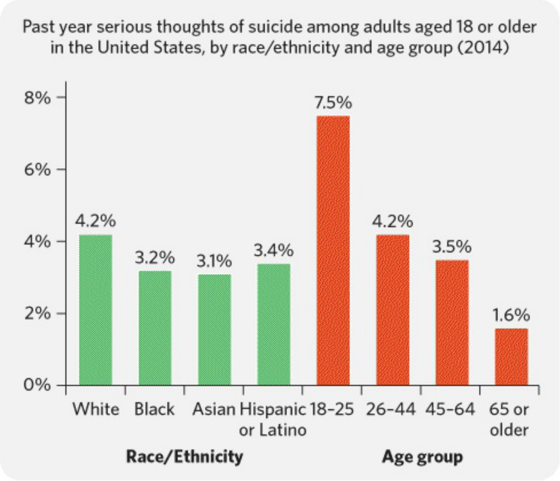


Figure 1.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.5 Bar graph showing the percent of American adults who had serious suicidal thoughts in 2014, for each of four ethnic groups and for each of four age groups.

STATS IN YOUR WORLD

CONVENIENT GRAPHS

The movie *An Inconvenient Truth* won the Oscar for Best Documentary in 2007. It was an unusual accomplishment for a movie featuring scientific evidence of global warming, supported by a detailed presentation of approximately 30 graphs and data tables.

APPLY YOUR KNOWLEDGE

1.3 Children's food choices. Does the presence of popular cartoon characters on food packages influence children's food choices? A study asked 40 young children (ages four to six) to taste two small pieces of Graham Crackers coming from a package with and a package without a popular cartoon character, and to indicate whether the two foods tasted the same or one tasted better. Unknown to the children, the crackers were the same both times. Here are the findings:[7]

Taste preference	Number of children	Percent
Without character	3	7.5
Taste the same	15	37.5
With character	22	55.0

a. Identify the individuals and the variable or variables in the study.
b. Present these data in a well-labeled bar graph.
c. Would it also be correct to present these data in a single pie chart? Explain your reasoning.
d. What do the data suggest about the influence of cartoon characters on Graham Cracker preference in young children?

1.4 More on children's food choices. The study in Exercise 1.3 also asked the 40 children to taste small pieces of gummy fruit snacks and baby carrots presented in packages with and in packages without a popular cartoon character. For each food type, the children indicated which of the two options they would prefer to eat for a snack. (Note that this is a different question from the one asked in Exercise 1.3.) The number and percent of children choosing the version with a cartoon on the package are displayed in the following table:

Food item	Number of children choosing the cartoon version	Percent choosing the cartoon version

Graham Crackers	35	87.5
Gummy fruit snacks	34	85.0
Baby carrots	29	72.5

a. Identify the individuals and the variable or variables in the study.

b. Make a well-labeled bar graph of the data.

c. Would it be correct to present these data in a single pie chart? Explain your reasoning.

d. What can you conclude from these findings?

QUANTITATIVE VARIABLES:

Histograms

Quantitative variables can take many values. The distribution of a variable tells us what values the variable takes and how often it takes these values. When there are more than just a few data points, a graph of the distribution is often easier to interpret if nearby values are grouped together. The most common graph of the distribution of one quantitative variable is a **histogram.**

EXAMPLE 1.5 Making a histogram: sharks

Stephen Frink Collection/Alamy

The great white shark, *Carcharodon carcharias*, is a large ocean predator at the top of the food chain. Here are the lengths in feet of 44 great whites:[8]

18.7	12.3	18.6	16.4	15.7	18.3	14.6	15.8	14.9	17.6	12.1
16.4	16.7	17.8	16.2	12.6	17.8	13.8	12.2	15.2	14.7	12.4
13.2	15.8	14.3	16.6	9.4	18.2	13.2	13.6	15.3	16.1	13.5
19.1	16.2	22.8	16.8	13.6	13.2	15.7	19.7	18.7	13.2	16.8

The *individuals* in this data set are individual sharks, and the *variable* is body length in feet. To make a histogram of the distribution of this variable, follow these steps:

STEP 1. Choose the classes. Divide the range of the data into classes of equal width. The data range from 9.4 to 22.8 feet, so we decide to use these classes:

9.0<individuals with body length≤11.011.0<individuals with body length≤13.0 ⋮ 21.0<individuals with body length≤23.0

$$\begin{aligned} 9.0 &< \text{individuals with body length} \leq 11.0 \\ 11.0 &< \text{individuals with body length} \leq 13.0 \\ &\vdots \\ 21.0 &< \text{individuals with body length} \leq 23.0 \end{aligned}$$

In this example, we chose to exclude the lower bound and include the upper bound in making the histogram classes. Choosing, instead, to include the lower bound and exclude the upper bound would also have been a valid option. What matters is specifying the classes precisely so that each individual falls into exactly one class. You can explain the nature of the class boundaries in the legend accompanying your histogram.

Based on the chosen class definition, the smallest shark, measuring 9.4 feet, falls into the first class; a shark measuring exactly 11.0 feet would still fall in that first class, but a shark measuring 11.1 feet would fall into the second class.

STEP 2. Count the individuals in each class. Here are the counts and corresponding percents:

Class	Count	Percent
9.1 to 11.0	1	2
11.1 to 13.0	5	11
13.1 to 15.0	12	27
15.1 to 17.0	15	34
17.1 to 19.0	8	18
19.1 to 21.0	2	5
21.1 to 23.0	1	2

Check that the counts add to 44, the number of individuals in the data (the 44 great white sharks) and that the percents add to 100 up to *roundoff error*. Here the rounded percents add to 99.

STEP 3. **Draw the histogram.** Graphs are typically created with statistical software rather than hand-drawn. The principle, however, is exactly the same for both approaches. Mark the scale for the variable whose distribution you are displaying on the horizontal axis—the shark body length in feet. The scale runs from 9 to 23 feet because that is the span of the classes we chose. The vertical axis contains either the scale of *counts* or the scale of *percents.* A histogram of percents rather than counts is convenient when we want to compare several distributions. It would, for instance, allow us to more easily compare the lengths of several shark species regardless of how many sharks were examined from each species. Each bar in the histogram represents a class. The base of the bar covers the class, and the bar height is the class percent. The horizontal axis represents a continuum of values (here body lengths), and therefore the histogram does not leave any horizontal space between the bars unless a class is empty (in which case the bar has height zero). Figure 1.6(a) is our histogram.

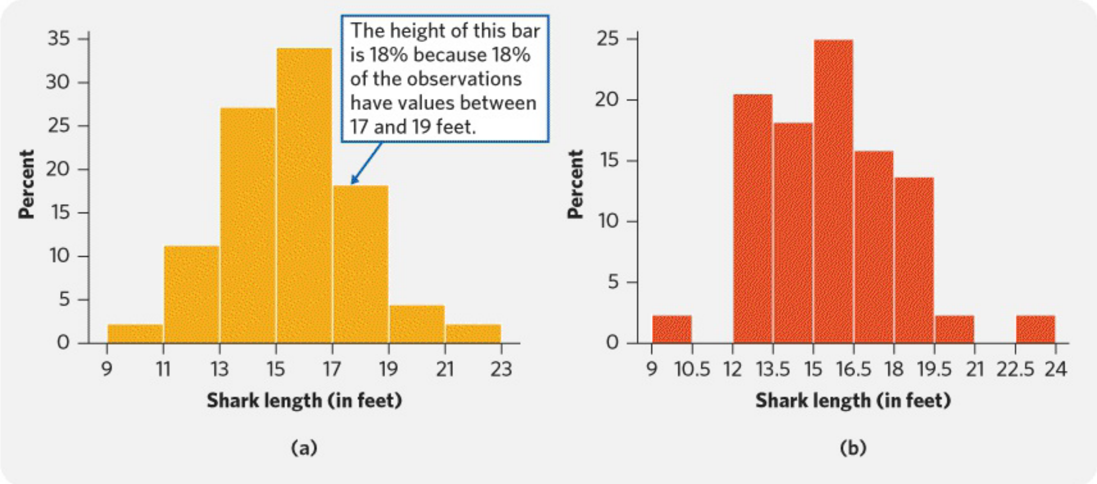


Figure 1.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.6 Histograms of the body length in feet for 44 great white sharks. (a) This histogram has 7 classes. (b) This histogram has 10 classes and shows more detail.

Although histograms resemble bar graphs in some aspects, their details and uses are very different. A histogram displays the distribution of one quantitative variable. The horizontal axis of a histogram is marked in the units of measurement for the variable. A

CAUTION

histogram for a continuous quantitative variable should be drawn with no extra space between consecutive classes, to indicate that all values of the variable are covered. In contrast, a bar graph compares different items that may be ordered any way we choose, as in Figure 1.3 (a, b). To emphasize this, a bar graph should separate the items being compared by leaving some blank space between the bars.

Our eyes respond to the *area* of the bars in a histogram.[9] Because the classes are all the same width, area is determined by height and all classes are fairly represented. There is no one "right" choice of the classes in a histogram. Too few classes will give a "skyscraper" graph, with all values in a few classes with tall bars. Too many classes will produce a "pancake" graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. You must use your judgment in choosing classes to display the shape. Statistics software will choose the classes for you (but *be aware that different software programs use different conventions for displaying the class boundaries*). The software's choice is usually a good one, but you can change it if you want. Figure 1.6(b) is a histogram of the shark data from Example 1.5 using class sizes of 1.5 feet rather than 2 feet. The histogram function in the *One-Variable Statistical Calculator* applet on the companion website allows you to change the number of classes by dragging with the mouse, so that it is easy to see how the choice of classes affects the histogram.

CAUTION

APPLET

APPLY YOUR KNOWLEDGE

1.5 Prescriptions of opioid pain relievers. Opioid pain relievers are prescribed at a higher rate in the United States than in any other nation, even though abuse of these medications can result in addiction and fatal overdoses. The CDC examined opioid pain reliever prescriptions in each state to find out how variable prescription rates are across the nation. Here are the 2012 state prescription rates, in number of prescriptions per 100 people, listed in increasing order:[10]

52.0 57.0 59.5 61.6 62.9 65.1 66.5 67.4 67.9 69.6

70.8	71.2	71.7	72.4	72.7	72.8	73.8	74.3	74.3	74.7
76.1	77.3	77.5	79.4	82.0	82.4	85.1	85.6	85.8	88.2
89.2	89.6	90.7	90.8	93.8	94.1	94.8	96.6	100.1	101.8
107.0	109.1	115.8	118.0	120.3	127.8	128.4	137.6	142.8	142.9

Make a histogram of the state opioid pain reliever prescription rates using classes of width 10 starting at 50.0 prescriptions per 100 people. (Make this histogram by hand even if you have software, to be sure you understand the process. You may want to compare your histogram with your software’s choice.)

1.6 Choosing classes in a histogram. The data set menu that accompanies the *One-Variable Statistical Calculator* applet includes a data set called “healing of skin wounds.” Choose these data, then click on the “Histogram” tab to see a histogram.

APPLET

a. How many classes does the applet choose to use? (You can click on the graph outside the bars to get a count of classes.)

b. Click on the graph and drag to the left. What is the smallest number of classes you can get? What are the lower and upper bounds of each class? (Click on each bar to find out.) Make a rough sketch of this histogram.

c. Click and drag to the right. What is the largest number of classes you can get? How many observations does the class with the highest count have?

d. The choice of classes changes the appearance of a histogram. Drag back and forth until you get the histogram you think best displays the distribution. How many classes did you use?

INTERPRETING HISTOGRAMS

Making a statistical graph is not an end in itself. The purpose of the graph is to help us understand the data. After you make a graph, always ask, "What do I see?" Once you have displayed a distribution, you can see its important features as follows.

EXAMINING A HISTOGRAM

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a histogram by its **shape, center,** and **spread.**

An important kind of deviation is an **outlier,** an individual value that falls outside the overall pattern.

We will learn how to describe center and spread numerically in Chapter 2. For now, we can describe the center of a distribution by its *midpoint*, the value at which roughly half the observations take smaller values and half take larger values. We can describe the spread of a distribution by giving the *smallest and largest values*.

EXAMPLE 1.6 Describing a distribution: sharks

Look again at the histogram in Figure 1.6(a).

SHAPE: The distribution is **unimodal.** That is, it has a **single peak,** which represents sharks with a body length between 15 and 17 feet. The distribution is also *symmetric*. In mathematics, the two sides of symmetric patterns are exact mirror images. Real data are almost never exactly symmetric. We are content to describe the histogram in Figure 1.6(a) as roughly symmetric.

CENTER: The counts in Example 1.5 show that 18 of the 44 sharks (41%) have a body length of 15 feet or less, but the count increases to 33 out of 44 (75%) for lengths of 17 feet or less. So the midpoint of the distribution is about 15 to 17 feet.

SPREAD: The spread is from 9.4 to 22.8 feet, but only 1 shark is shorter than 11 feet and only 1 is more than 21 feet long.

OUTLIERS: In Figure 1.6(a), the observations greater than 21 feet or less than 11 feet are part of the continuous range of body lengths and do not stand apart from the overall distribution. This histogram, with only 7 classes, hides some of the detail in the distribution. Figure 1.6(b) is a histogram of the same data with 10 classes instead. It reveals that the shortest and longest sharks, at 9.4 and 22.8 feet, respectively, do stand apart from the other observations.

Once you have spotted possible outliers, look for an explanation. Some outliers are due to mistakes, such as typing 19.4 as 9.4. Other outliers point to the special nature of some observations. The smallest shark might be a juvenile exhibiting some adult features, for instance. An outlier could also simply be an unusual but perfectly legitimate observation. For instance, the largest shark could be just that: an unusually large shark. Human height, for instance, has a somewhat homogeneous distribution within a gender and ethnicity, but some individuals, such as the famous basketball player Shaquille O’Neal, clearly stand out from the norm.

CAUTION

Comparing Figures 1.6(a) and (b) reminds us that *the choice of classes in a histogram can influence the appearance of a distribution.* Both histograms portray a symmetric distribution with one peak, but only Figure 1.6(b) shows the mild outliers.

When you describe a distribution, concentrate on the main features. Look for major peaks, not for minor ups and downs in the bars of the histogram. Look for clear outliers, not just for the smallest and largest observations. Look for rough *symmetry* or clear *skewness*.

SYMMETRIC AND SKEWED DISTRIBUTIONS

A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other.

A distribution is **skewed to the right,** or positively skewed, if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is **skewed to the left,** or negatively skewed, if the left side of the histogram extends much farther out than the right side.

Here are more examples of describing the overall pattern of a histogram.

EXAMPLE 1.7 A pertussis epidemic

Pertussis, also known as whooping cough, is a highly contagious respiratory disease characterized by uncontrollable, violent coughing. Pertussis can be fatal, especially among infants too young to be vaccinated. As mentioned in Example 1.3, California experienced an unusually large epidemic of pertussis in 2014, with 11,203 cases reported in that year. Here are the rates of pertussis cases per 100,000 persons reported in each of the 59 California counties for all of 2014:[11]

0.0	0.0	0.0	0.0	0.0	0.0	2.8	3.4	3.5	4.7
5.2	5.6	7.2	8.1	8.2	9.5	10.0	11.3	13.6	14.5
14.6	14.9	15.3	15.5	16.6	17.4	17.4	17.4	18.5	18.8
19.2	19.6	20.3	20.8	25.1	27.8	29.5	29.9	30.0	30.7
30.7	30.9	33.4	34.0	37.2	37.3	40.9	41.8	44.1	48.6
59.9	60.4	63.4	64.7	71.3	99.2	107.1	109.8	143.0	

The histogram in Figure 1.7 indicates that the distribution of pertussis rates is *single-peaked* and obviously strongly *skewed to the right.* Most counties had a relatively small rate, between 0 and 40 cases per 100,000 persons, and they make up the peak of the distribution (a single peak can span several classes, since the choice of classes is somewhat arbitrary). However, some counties has higher rates, so that the graph extends substantially to the right of its peak. This is not surprising considering that epidemics tend to cluster geographically. Here, the center (half above, half below) lies in the first class (0 to 20 cases per 100,000 persons). The pertussis rates range from 0 to 143 cases per 100,000 persons, though the county with the largest rate (Sonoma County) is a clear *outlier*.

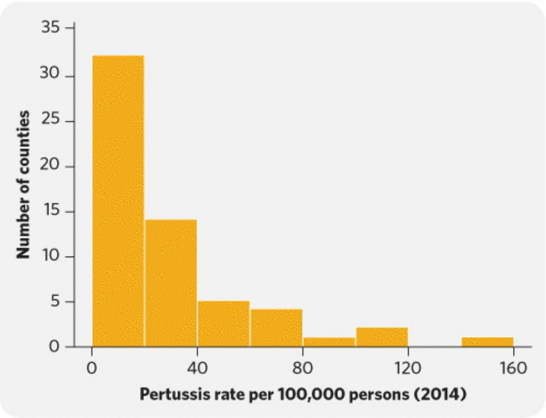


Figure 1.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.7 Histogram of 2014 rates of pertussis (in number of cases per 100,000 persons) in the 59 California counties. Classes include the lower-bound value but exclude the upper-bound value.

STATS IN YOUR WORLD

THE VITAL FEW

Skewed distributions can show us where to concentrate our efforts. Ten percent of the cars on the road account for half of all carbon dioxide emissions. A histogram of CO_2 emissions would show many cars with small or moderate values and a few with very high values. Cleaning up or replacing these high-emission cars would reduce pollution at a cost much lower than that of programs aimed at all cars. Statisticians who work at improving quality in industry often rely on this principle: Distinguish “the vital few” from “the trivial many.”

EXAMPLE 1.8 Lyme disease

Lyme disease is a bacterial infection spread through the bite of an infected blacklegged tick. Left untreated, it can cause lifelong complications. Figure 1.8 displays data about the age in years of 241,931 individuals diagnosed with Lyme disease in the United States between 1992 and 2006.[12] We can't call this irregular distribution either symmetric or skewed. The major feature of the overall pattern is the presence of two main peaks; that is, the data have a **bimodal** distribution corresponding to two **clusters** of individuals—children and adults.

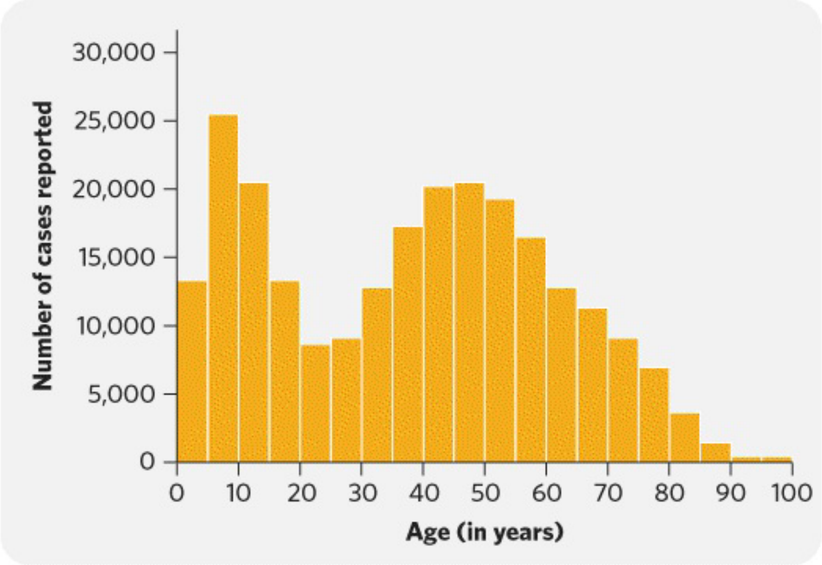


Figure 1.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.8 Histogram of patient age in years for 241,931 cases of Lyme disease reported in the United States between 1992 and 2006. Notice the two separate peaks around 10 years and 45 years.

Clusters suggest that several types of individuals are mixed in the data set. The first cluster has a clear peak around 5 to 15 years, whereas the peak of the second cluster is at 45 to 50 years. Because ticks are found outdoors in woody or grassy areas, these two clusters might reflect the outdoor activities of children and the adults accompanying them.

Giving a single center and spread for this distribution would be misleading, because the data suggest two age groups. It would be better to describe the two groups separately.

The overall shape of a distribution provides important information about a variable. Some variables have distributions with predictable shapes. Many biological measurements on specimens from the same species and sex—lengths of bird bills, heights of young women—have symmetric distributions. In contrast, survival times—patients after an organ transplant, lab animals following an experimental inoculation—have distributions that are typically strongly skewed to the right. Many distributions have irregular shapes that are neither symmetric nor skewed. Some data show other patterns, such as the clusters in Figure 1.8. *Do not try to artificially manipulate the classes of a histogram so that the data appear more symmetrical or more regular.* Instead, accept that not all data have a distribution that follows a "neat" pattern, even if you could obtain a larger data set. Use your eyes, describe what you see, and then try to explain it.

APPLY YOUR KNOWLEDGE

1.7 Prescriptions of opioid pain relievers. In Exercise 1.5, you made a histogram of the state prescription rates for opioid pain relievers. Would you say that the distribution is single-peaked or multiple-peaked? Is it roughly symmetric or skewed?

1.8 Age of onset for anorexia nervosa. Anorexia nervosa is an eating disorder characterized, in part, by the fear of becoming fat and by voluntary starvation; it often leads to death from suicide or organ failure. A Canadian study surveyed 691 adolescent girls diagnosed with anorexia nervosa and their families.[13] Figure 1.9 is a histogram of the distribution of age at onset of anorexia for the 691 girls in the study (note that, if the onset occurred on the 12th birthday, for example, the onset is reported as 12.01 for the purpose of the histogram, to be coherent with customary age descriptions). Describe the shape of this distribution. Within which class does the midpoint of the distribution lie?

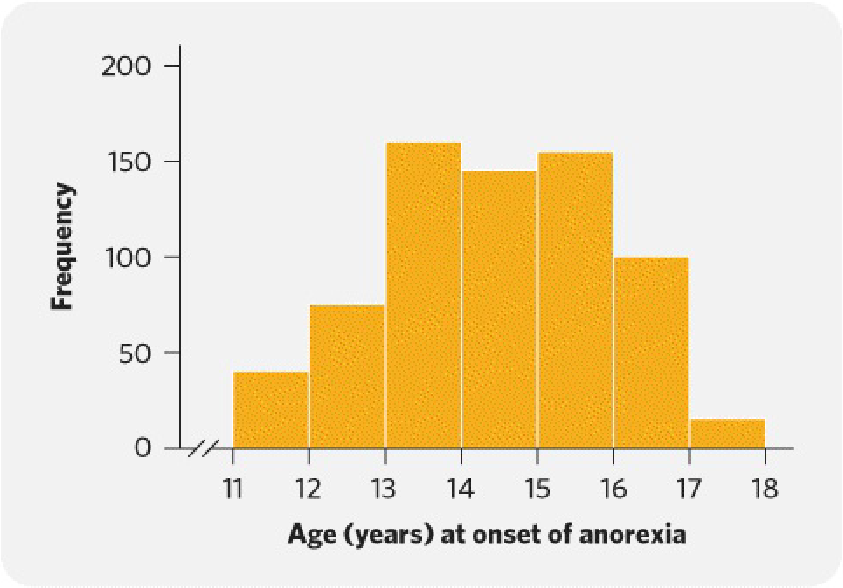


Figure 1.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.9 Histogram of age (in years) at onset of anorexia nervosa for 691 Canadian girls diagnosed with the disorder. The first class includes 11-year-old girls but excludes 12-year-olds.

PhotoStock-Israel/Getty Images

QUANTITATIVE VARIABLES:

Dotplots

Histograms are not the only graphical display of quantitative distributions. *Dotplots* are also commonly used to diplay the distribution of quantitative data, especially for small data sets. They have the added advantage of displaying the **raw data;** that is, they show each one of the values in the data set.

DOTPLOT

To make a **dotplot:**

1. Sort the data set and plot each observation according to its numerical value along a labeled scaled axis.
2. Identical observations are typically stacked.

EXAMPLE 1.9 Making a dotplot: sharks

In Example 1.5 we saw how the body lengths of 44 great white sharks can be plotted in a histogram to more easily examine their distribution. Any raw data that can be displayed in a histogram can also be displayed in a dotplot, and vice versa. To make a dotplot of the shark data, create a one-dimensional graph with shark length on the axis (in feet). The axis may be horizontal or vertical, as desired, because there is only one axis. Figure 1.10(a) shows one dotplot of this data set. Here, data points with the same or very similar numerical values are shown as stacked dots. For example, two sharks had a body length of 18.7 feet and they are shown as two stacked dots. The sharks with body lengths of 12.1 and 12.2 feet are also shown as two stacked dots.

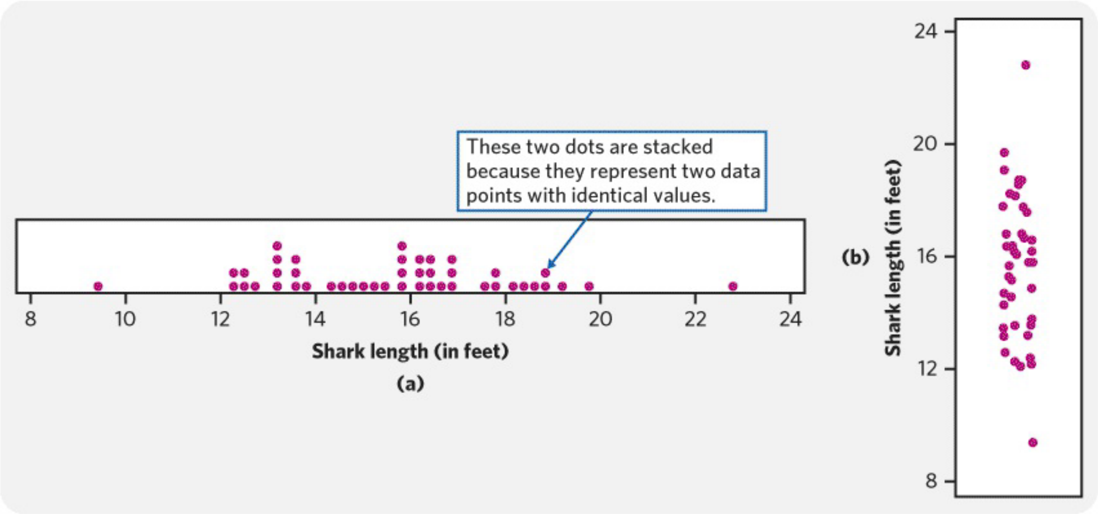


Figure 1.10

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.10 Dotplot of body length (in feet) for 44 great white sharks. (a) This dotplot is scaled horizontally, with similar values shown as stacked dots. (b) This dotplot is scaled vertically, with similar values distinguished by adding jitter between the dots.

Like a histogram, the dotplot shows the shape, center, and spread of the distribution as well as potential outliers. However, because the dotplot is one-dimensional, the distribution's shape and center are indicated by the density of dots rather than the height of the histogram bar. Notice how the dots in Figure 1.10(a) are fairly evenly spread in a roughly symmetric fashion, and that the two mild outliers stand apart from the rest of the data.

Dotplots with neatly stacked dots are easy to draw by hand for reasonably small data sets. When using statistical software, you will typically have the option of adding jitter between dots with similar numerical values so that they can be plotted at their exact location on the scaled axis while still being easily distinguishable, as shown in Figure 1.10(b).

APPLY YOUR KNOWLEDGE

1.9 California pertussis epidemic. In Example 1.7 we used a histogram to display the rates of pertussis cases per 100,000 persons in each of the 59 California counties for all of 2014. Create a dotplot of the data, by hand or using software. Compare your dotplot to the histogram in Figure 1.7 and explain the differences and similarities between the two plots.

1.10 Glucose levels. People with diabetes must monitor and control their blood glucose level. The goal is to maintain “fasting plasma glucose” between about 90 and 130 milligrams per deciliter (mg/dl). A study planned to compare the effectiveness of group instruction versus individual instruction on diabetes control. Here are the fasting plasma glucose levels for 18 diabetic patients who were given group instruction and 16 diabetic patients who were given individual instruction, five months after the end of instruction:[14]

Group instruction									
141	158	112	153	134	95	96	78	148	172
200	271	103	172	359	145	147	255		

Individual instruction									
128	195	188	158	227	198	163	164	159	128
283	226	223	221	220	160				

a. Make one dotplot containing both data sets. Software produces comparative dotplots easily. For handmade graphs, create a dotplot with one horizontal axis but two parallel lines, one over the other. Use one line to place the group instruction dots and the other line to place the individual instruction dots.

b. Describe the main features of each distribution. Are there any outliers?

c. How well is each group, as a whole, achieving the goal of controlling glucose level? How do the two groups differ with respect to this goal?

STATS IN YOUR WORLD

FLORENCE NIGHTINGALE

Florence Nightingale (1820–1910) dedicated her life to promoting and improving the field of nursing. She recorded data and displayed her findings in graphs that policymakers could understand. During the Crimean war, she showed that more soldiers died because of poor hospital conditions than from battle wounds. She also documented how casualties drastically dropped after sanitary conditions were improved. Nightingale was the first woman elected to the Royal Statistical Society, honoring her outstanding contributions.

TIME PLOTS

Many variables are measured at intervals over time. We might, for example, measure the height of a growing child or water precipitation at the end of each month. In these examples, our main interest is in change over time. To display change over time, make a *time plot*.

TIME PLOT

A **time plot** of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

EXAMPLE 1.10 Water salinity in the Shark River

The Shark River runs through the southwestern portion of Everglades National Park in Florida and flows into the Gulf of Mexico. The U.S. Geological Survey closely monitors this important natural habitat. Figure 1.11 is a time plot of water salinity at the Gunboat Island station on the Shark River over a seven-day period in the fall of 2009. Salinity was recorded every 15 minutes from 12 A.M. (midnight) of September 30 to 12 A.M. (midnight) of October 7.[15]

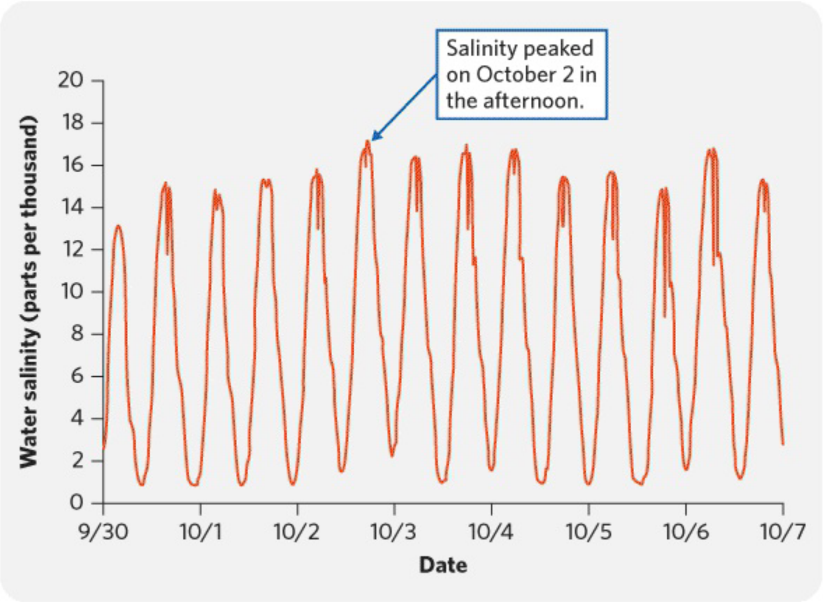


Figure 1.11

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.11 Time plot of Shark River water salinity in Everglades National Park over a seven-day period in 2009. The daily cycles reflect the influence of the tides from the Gulf of Mexico, into which the Shark River discharges.

When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. Figure 1.11 shows strong **cycles,** representing regular up-and-down movements in water salinity. The cycles show the effects of the ocean tides on the salinity of the Shark River water. Water salinity in the river is highest twice a day, every day, coinciding with the high tides. This cyclical pattern is very clear and consistent, despite some variations in salinity from one day to the next.

EXAMPLE 1.11 Global warming

Global warming is an important planet-wide ecological issue that has been hotly debated. Table 1.1 lists the annual global temperature anomaly (in degrees Celsius) from 1880 to 2012 based on data from recording stations around the world.[16] Individual annual temperature anomalies are computed locally by comparing the local annual sea surface temperature average with the local temperature reference (the 1951–1980 average). Both graphs in Figure 1.12 describe these data.

TABLE 1.1 **Annual global temperature anomaly (degrees Celsius)**

Year	Anomaly	Year	Anomaly	Year	Anomaly	Year	Anomaly	Year	Anom
1880	−0.25	1903	−0.31	1926	−0.01	1949	−0.06	1972	0
1881	−0.19	1904	−0.34	1927	−0.13	1950	−0.15	1973	0.1
1882	−0.22	1905	−0.24	1928	−0.11	1951	−0.04	1974	−0.
1883	−0.23	1906	−0.2	1929	−0.24	1952	0.03	1975	−0.
1884	−0.3	1907	−0.38	1930	−0.06	1953	0.11	1976	−0.1
1885	−0.3	1908	−0.34	1931	−0.01	1954	−0.1	1977	0.1
1886	−0.25	1909	−0.35	1932	−0.06	1955	−0.1	1978	0.0
1887	−0.35	1910	−0.33	1933	−0.17	1956	−0.17	1979	0.0
1888	−0.26	1911	−0.33	1934	−0.05	1957	0.07	1980	0.1
1889	−0.15	1912	−0.34	1935	−0.1	1958	0.08	1981	0.2
1890	−0.36	1913	−0.32	1936	−0.04	1959	0.06	1982	0.0
1891	−0.28	1914	−0.15	1937	0.08	1960	−0.01	1983	0.2
1892	−0.32	1915	−0.09	1938	0.11	1961	0.08	1984	0.0
1893	−0.31	1916	−0.3	1939	0.03	1962	0.04	1985	0.0
1894	−0.33	1917	−0.4	1940	0.05	1963	0.08	1986	0.1
1895	−0.27	1918	−0.32	1941	0.11	1964	−0.21	1987	0.2
1896	−0.16	1919	−0.2	1942	0.04	1965	−0.11	1988	0.3
1897	−0.12	1920	−0.18	1943	0.1	1966	−0.03	1989	0.1
1898	−0.24	1921	−0.13	1944	0.2	1967	0	1990	0.3
1899	−0.17	1922	−0.24	1945	0.07	1968	−0.04	1991	0.3
1900	−0.1	1923	−0.2	1946	−0.04	1969	0.08	1992	0.1
1901	−0.15	1924	−0.21	1947	0.01	1970	0.03	1993	0.1
1902	−0.27	1925	−0.16	1948	−0.04	1971	−0.1	1994	0.2

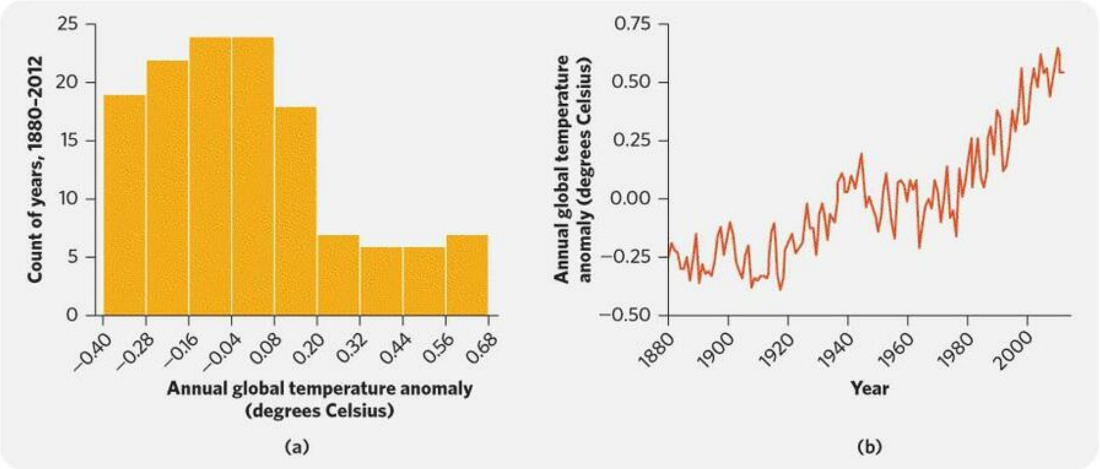


Figure 1.12

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.12 Histogram (a) and time plot (b) of the annual global temperature anomaly over the 133 years from 1880 to 2012. Data are from Table 1.1. Anomalies are relative to the 1951–1980 reference period.

© Eric Lefranc/Biosphoto

The histogram in Figure 1.12(a) shows the distribution of annual global temperature anomalies. The histogram is unimodal (single-peaked) and skewed to the right, with a center located at approximately −0.04 to 0.08 degrees Celsius. We might think that these data show just chance year-to-year fluctuations in global temperature anomalies, with a small proportion of larger anomalies.

Figure 1.12(b) is a time plot of the same data. For example, the first point lies above 1880 on the “Year” scale at height −0.25, the global temperature anomaly for 1880. This time plot tells a more interesting story than the histogram. There is a great deal of year-to-year variation, but there is also a clear increasing **trend** over time. That is, there is a long-term rise in annual global temperature anomalies. A trend in a time plot is a long-term upward or downward movement over time. The trend in the temperature data reflects a climate change: Global temperatures have been rising fairly steadily for more than a century, despite substantial year-to-year variations.

Histograms and time plots give different kinds of information about a variable. The time plot in Figure 1.12(b) presents **time series data** that show the change in annual global temperature anomalies over time. A histogram displays the distribution of data, such as the temperature anomalies, regardless of time.

APPLY YOUR KNOWLEDGE

1.11 Opioid pain relievers. In Exercise 1.5, we looked at state prescription rates for opioid pain relievers. Here are the number of deaths from prescription opioid pain relievers in the United States in the years 2000 to 2014:[17]

Year	2000	2001	2002	2003	2004	2005	2006	2007
Number of deaths	4,400	5,528	7,456	8,517	9,857	10,928	13,723	14,408
Year	2008	2009	2010	2011	2012	2013	2014	
Number of deaths	14,800	15,597	16,651	16,917	16,007	16,235	18,893	

Make a time plot of these data and describe the trend that your plot shows. What does the trend emphasize?

Stephen Frink Collection/Depositphotos

1.12 Atmospheric CO_2 levels. Elevated atmospheric CO_2 levels have been linked to global warming. The Mauna Loa Observatory in Hawaii has the longest record of direct measurements of CO_2 in the atmosphere, going back to 1958. Figure 1.13 shows the trends in monthly atmospheric CO_2 levels over Hawaii between March 1958 and January 2013.[18] Describe the patterns in these data. What do the data suggest about controlling our carbon footprint?

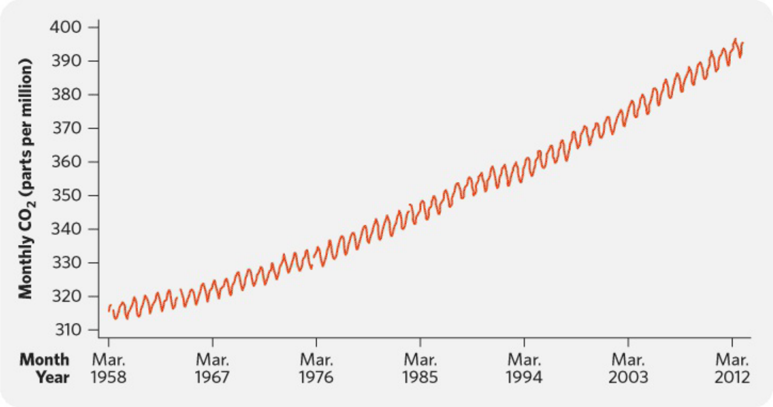


Figure 1.13
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.13 Time plot of monthly atmospheric CO_2 levels recorded at Mauna Loa Observatory, from March 1958 to January 2013.

DISCUSSION (Mis)adventures in data entry

Data are at the center of statistics, but how do you record data so that they can be analyzed easily? And what are some issues you should be concerned about? This discussion provides a few pointers. Failure to pay attention to these issues can lead to disastrous results, ranging from having to redo your whole analysis, to needing to collect new data or, worse yet, publishing incorrect findings and having to rescind them later.

Keeping detailed records

Scientific inquiries must be documented and you will be required to keep clear, detailed records of your findings. This is a necessary safeguard against dishonest conduct, but it is also a great help in data entry and proofing. To avoid accusations of data falsification or disputes over who discovered something first, using bound notebooks with page numbers and written dates to record data is good practice. You might also keep photographs of your findings or actual specimens. The discussion in Chapter 2 on dealing with outliers highlights how the first step after identifying an outlier is to check your data records and stored evidence to see if the outlier could have been a typographical or other simple data entry error.

Along with the actual data, you should record anything relevant to how the data were obtained, measured, or computed. For example, are values of weight recorded in grams, kilograms, ounces, or pounds, and are they self-reported (common in telephone surveys) or actually measured with a scale? If you computed body mass index from values of weight and height, exactly how did you perform that computation and did it involve rounding numerical values? This step will help you communicate your findings after your analysis is done, and it will be very useful if you ever find puzzling values during the course of your analysis.

Organizing data for use with statistical software

Whether you write your scientific findings on paper or record them electronically, at some point you will want to use statistical software for graphing and statistical analysis. Spreadsheet programs such as Microsoft Excel or LibreOffice Calc are commonly used for this purpose because they are widely available and because they allow data, notes, and even pictures to be saved into one convenient space. You can think of this approach as the electronic equivalent of your paper-based notebook. Proper statistical analysis, however, often requires that your data be organized in very specific ways.

Statistical software packages differ quite a bit in terms of which data format they require, sometimes even requiring different formats for different analyses. Let's just say that no one in the software business seems to have given serious consideration to the painful experiences of the users…Statisticians, however, generally agree that the best way to record your data is to use one row for each individual and one column for each variable. The table in Example 1.1 follows this format: Each tree shrew is given its own row, and every variable recorded about these tree shrews is given its own column (including a unique identifier, here the animal's given name).

Why is the "one individual, one row" approach optimal for data analysis? Let's consider the alternatives. (1) The researchers in Example 1.1 probably had information about each animal's sex and age before they obtained the data on paw usage, so it might have been tempting to use a different spreadsheet for the animals' personal information and for the experimental findings. However, this would make statistical analysis a lot more difficult if they wanted to, say, compare the male and female shrews. (2) Alternatively, the researchers could have recorded their findings on paw usage separately for the male and the female shrews, creating two different data tables. The challenge, then, would be performing an analysis using all shrews (regardless of sex).

When entering categorical values, be sure to be consistent with your choice of naming convention. Statistical software will consider "m" and "male" two different things, even if you understand that both mean the shrew is male. Some software packages won't accept text at all, and instead require that you use a numerical code, such as 0 for male and 1 for female. In such a case it would be better to use "Female" rather than

“Sex” as the variable name and to record it as 0 for no and 1 for yes, which would be easier to interpret.

Lastly, there will be times when you have to deal with missing data. Maybe a tree shrew died before you collected all the data or it refused to perform one of the food tasks in your study—or perhaps you forgot to write that particular piece of data in your notebook…These things happen. But how do you handle it in your electronic data records? The answer, unfortunately, depends on which statistical software you use. Some applications mark missing data with an asterisk, others with a dot, the letters “NA,” or a blank space. This can create serious complications if you need to transfer your data analysis from one software application to another. Statisticians sometimes recommend using a special code for missing data, such as 0 or −99 (using different codes for different causes for the missing data). This is a great strategy if the person recording the data is also the person performing the data analysis, but it can be disastrous if the person running the data analysis is not aware of this convention.

Checking for obvious errors, inconsistencies, and missing values

Data entry does not end with the creation of an electronic file of data values. You need to make sure your data are correct. Even data entry professionals are not 100% accurate. Every now and then, you can expect to face problems like spelling differences (as in the “m” and “male” mentioned earlier for categorical data), typos (notice how the values 9 and 0 are right next to each other on the keyboard’s top row, and the values 9 and 6 on the numerical keypad), or a skipped row (this mistake is almost too easy to make). The real issue is catching them.

All of your data values should be realistic or at least plausible biologically. In an online class survey asking for height in inches, a student wrote “6.0” as the answer—this is definitely *not* realistic. If you do not catch this obvious error, all your analyses will be affected. In Chapters 2, 3, and 4, we discuss at length how outliers impact various statistical computations.

The best way to catch a mistake like an impossible height value is to plot your data. Stemplots, dotplots, and histograms are excellent ways to *see* unusual or implausible values. Numerical summaries and boxplots (described in Chapter 2) work, too. Time plots for data collected over a period of time are also useful for revealing events that influenced the data collection process (for example, data values becoming systematically larger after switching to a new measuring instrument or when taken by a different person). If your data are categorical, obtain a summary frequency table. You would see, for example, four values for Sex in such a table if you carelessly recorded sex as either m or male, and f or female.

Missing data can create substantial problems, especially if they go unnoticed. The publicly available Pima Indians diabetes data set, for example, contains biological data for 768 female members of the Pima Indian tribe. The data set is listed as having no missing data values. Yet, careful graphical exploration of the data reveals large numbers of zeros that are not biologically plausible (for example, for values of blood pressure or skinfold thickness). Perhaps even more troubling, it appears that many published analyses using this data set failed to identify the problem and instead treated such zeros as actual biological values.[19]

Don't make the same mistake. If you do not plot your data and carefully check them before running further statistical analyses, you could waste a lot of time and energy and, much worse, come to erroneous conclusions. Be smart. Plot your data!

APPLY YOUR KNOWLEDGE

1.13 Student data. Consider gathering some basic information about your classmates, perhaps during lecture or discussion.

a. You may be interested in the distribution of student heights. Which type of variable is student height? Explain how you would collect the data if the class had 30 or 40 students. Which kind of graph would you use to display the data?
b. You may also be interested in the class gender breakdown. Which type of variable is student gender? Explain how you would collect the data if the class had 30 or 40 students. Which

kind of graph would you use to display the data?

c. We know that, overall, men tend to be taller than women. Therefore, it would be a good idea to graph and interpret the student height data separately for male and female students. Would the way you collected the data in parts a and b allow you to do this? How would you need to collect and organize the data to allow a comparison of heights for both genders? Which part of the Discussion on page 26 addresses this issue?

1.14 Temperatures records. NASA's Goddard Institute for Space Studies (GISS) allows you to download historical temperature records for locations worldwide. Go to http://data.giss.nasa.gov/gistemp/station_data, scroll down to "Download Station Data," and type "Los Angeles." (You may be interested in looking up your own location as well.) A new page appears with a link to the "Los Angeles California" station data. Click on this link to the next page, which should contain a time plot of the annual mean temperature (in degrees Celsius) for Los Angeles since 1880 and a link to "Download monthly data as text." (If this URL isn't working, you can find the data set on the companion website for this textbook, ex01-14.)

a. Open this data file in your preferred statistical software and create a dotplot of the meteorological annual mean temperatures (last column, "metANN"). Describe the distribution. Is there anything suspicious about the data? What could explain it?

b. Go back to the GISS website and look at the time plot of annual mean temperature. Notice that the line is interrupted in some parts of the graph. With this new insight, what do you think is the most likely explanation for the suspicious pattern in your dotplot? Which part of the Discussion on page 26 addresses this issue?

c. Now clean up the data so that only true temperature values will be plotted. You can do this by editing either the original text file or the data in your statistical software. Take the time to consider how your specific software handles missing data (if an empty cell won't work, see your software's help function).

d. Create both a dotplot and a time plot of the cleaned-up annual mean temperatures. Interpret your graphs and conclude in context.

CHAPTER 1 SUMMARY

- A data set contains information on a number of **individuals.** Individuals may be people, animals, or things. For each individual, the data give values for one or more **variables.** A variable describes some characteristic of an individual, such as a person's height, sex, or age.
- Some variables are **categorical** and others are **quantitative.** A categorical variable places each individual into a category, such as male or female. A quantitative variable has numerical values that measure some characteristic of each individual, such as height in centimeters or age in years.
- **Exploratory data analysis** uses graphs and numerical summaries to describe the variables in a data set and the relations among them.
- After you understand the background of your data (individuals, variables, units of measurement), the first thing to do almost always is **plot your data.**
- The **distribution** of a variable describes what values the variable takes and how often it takes these values. **Pie charts** and **bar graphs** display the distribution of a categorical variable. Bar graphs can also compare any set of quantities measured in the same units. **Histograms** and **dotplots** display the distribution of a quantitative variable.
- When examining any graph, look for an **overall pattern** and for notable **deviations** from the pattern.
- **Shape, center, and spread** describe the overall pattern of the distribution of a quantitative variable. Some distributions have simple shapes, such as **symmetric** or **skewed.** Not all distributions have a simple overall shape. Describing the shape of a distribution when there are few observations can be particularly challenging.
- **Outliers** are observations that lie outside the overall pattern of a distribution. Always look for outliers and try to explain them.
- When observations on a variable are taken over time, make a **time plot**

that graphs time horizontally and the values of the variable vertically. A time plot can reveal **trends, cycles,** or other changes over time.

THIS CHAPTER IN CONTEXT

Practical statistics is the science of extracting meaning out of validly collected data. We will see in Chapters 6 and 7 that the way we collect data drastically affects the conclusions that we may draw. Some studies examine data from an entire population of interest, as in Example 1.3 on all reported cases of infectious diseases for the state of California in 2014. Other studies select a sample from a given population, as did the study in Example 1.1 about the paw usage preferences of 36 tree shews. The ultimate objective, though, is to draw conclusions about the wider population (for example, all tree shrews), a process called statistical inference that we will describe starting in Chapter 14.

Regardless of the ultimate objective, understanding data starts with exploratory data analysis: the use of graphs and numerical summaries to reveal patterns. This may be done purely for a descriptive purpose—as highlighted in Part I of this book. Alternatively, it may be done to check whether the data we have are suitable for a specific inference procedure—something we will cover in Parts III and IV. In addition, the discussion box on page 26 illustrates how graphs and numerical summaries are useful tools for finding inconsistencies in the data (typos and other errors) before any real analysis can begin.

In this chapter we showed how categorical data can be summarized graphically using pie charts and bar graphs and how the distribution of a quantitative variable can be displayed with histograms and dotplots—or, for time series, with time plots. These are the simplest and most commonly used types of graphs for inspecting one variable at a time. In Chapters 3 through 5 we will also discuss how graphs can be used to examine the relationship between two quantitative variables or two categorical variables. More complex graphs are sometimes used in the life sciences, such as graphs mapping data to a geographical location to depict patterns of epidemic spread, animal migration, or climate change. Their study is beyond the scope of this introductory textbook.

CHECK YOUR SKILLS

1.15 A study of a very large number of pregnant women in Arkansas reports that the women gained, on average, 14 pounds during their pregnancy and that 18% of the women smoked.[20] Which of the following is *not* a variable in this study?

a. Pregnancy status
b. Smoking status
c. Weight gain

1.16 The two variables in the Arkansas study are

a. both categorical variables.
b. both quantitative variables.
c. one categorical variable and one quantitative variable.

The Statistical Abstract of the United States, *prepared by the Census Bureau, provides the number of single-organ transplants for the year 2010, by organ. The next two exercises are based on the following table:*

Heart	2,333
Lung	1,770
Liver	6,291
Kidney	16,898
Pancreas	350
Intestine	151

1.17 The data on single-organ transplants can be displayed in

a. a pie chart but not a bar graph.
b. a bar graph but not a pie chart.
c. either a pie chart or a bar graph.

1.18 Kidney transplants represented what percent of single-organ transplants in 2010?

a. Nearly 61%

b. One-sixth (nearly 17%)

c. This percent cannot be calculated from the information provided in the table.

Figure 1.14 shows the percent of adults in the world who are overweight or obese, by type of country of residence based on that country's income level.[21] *The following two exercises are based on this figure.*

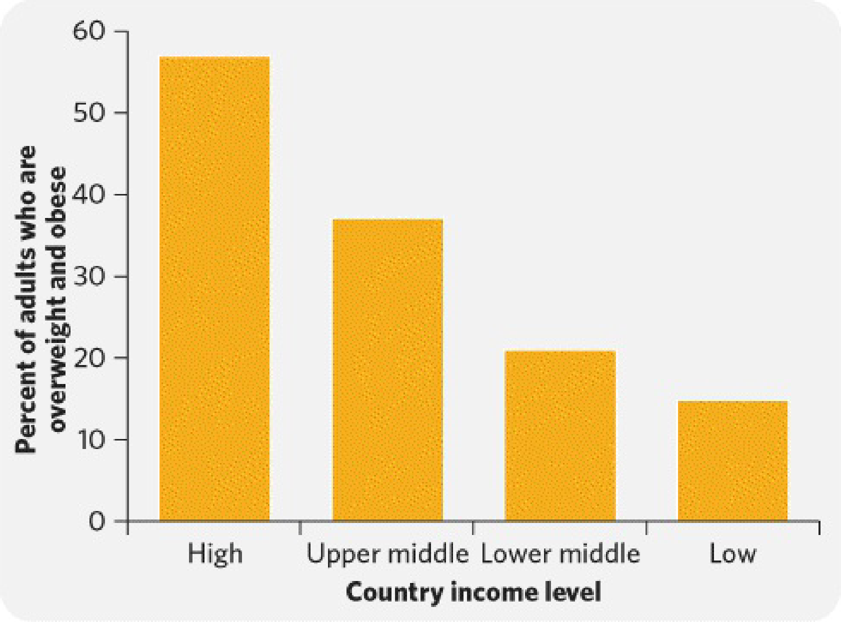


Figure 1.14
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.14 Percent of adults who are overweight or obese in the world, by country income level.

1.19 The graph in Figure 1.14 is

a. a bar graph that can be made into one pie chart.

b. a bar graph that cannot be made into one pie chart.

c. a histogram with a clear right skew.

1.20 Which of the following conclusions can be reached from Figure 1.14?

a. The majority of adults who are overweight and obese live in high-income countries.

b. The majority of adults who live in high-income countries are overweight and obese.

c. Both conclusions are correct.

Figure 1.15 is a histogram of the takeoff angles of 51 videotaped jumps of adult hedgehog fleas, Archaeopsyllus erinacei.[22] *The following two exercises are based on this histogram.*

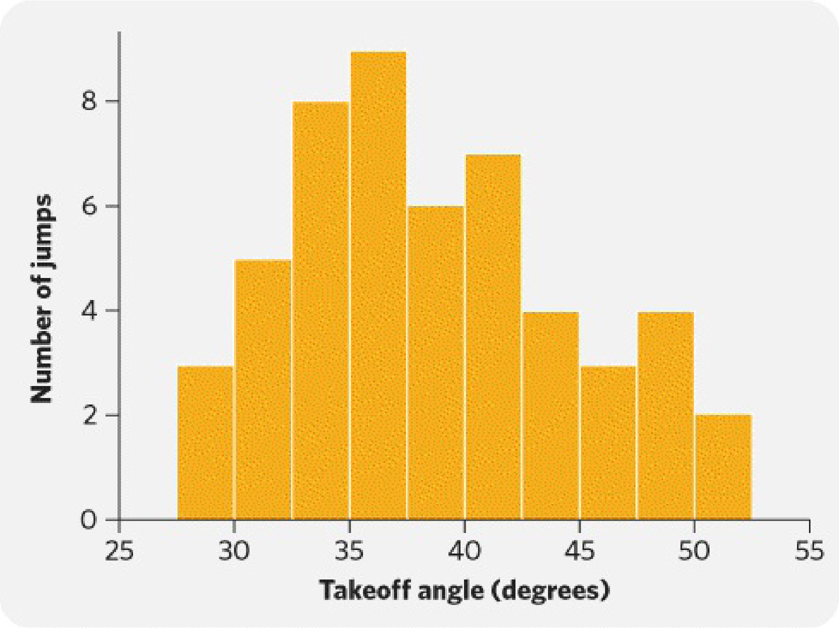


Figure 1.15
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.15 Histogram of the takeoff angles (measured in degrees) of 51 flea jumps.

1.21 What percent of jumps have a takeoff angle of 35 degrees or less?

a. 8%
b. 16%
c. 31%

1.22 The shape of the distribution of takeoff angles in Figure 1.15 is

a. skewed to the right.
b. roughly symmetric.
c. skewed to the left.

1.23 The 2010 U.S. census reported the percent of individuals younger than age 18 in each of the 50 states and the District of Columbia. The data are shown in a doplot in Figure 1.16. This distribution is

a. single-peaked without outliers.
b. single-peaked with two outliers.
c. multiple-peaked with two outliers.

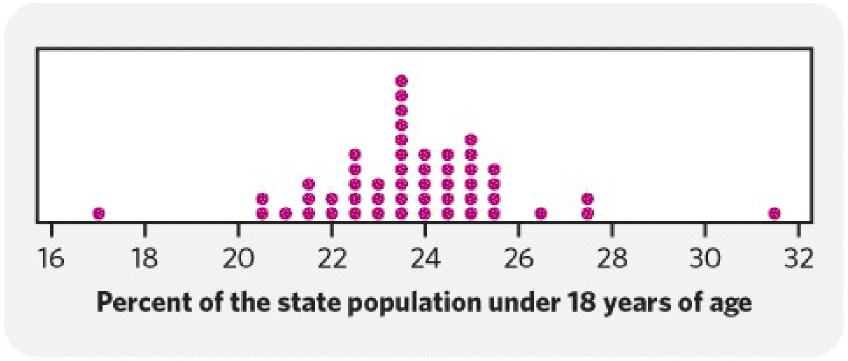


Figure 1.16
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.16 Dotplot of the percent of population younger than 18 years of age for the 50 states and D.C.

1.24 The shape of the distribution in the previous exercise is

a. strongly skewed to the right.
b. roughly symmetric.
c. strongly skewed to the left.

CHAPTER 1 EXERCISES

1.25 Endangered species. Bald eagles are an endangered bird species suffering from loss of habitat and pesticide contamination of rivers. A field biologist studying the reproduction of bald eagles records data for the following variables. Which of these variables are categorical, and which are quantitative?

a. Number of eggs laid
b. Incubation period (in days)
c. Parental care (mostly mother, mostly father, both parents)
d. Nest size (in centimeters)
e. Presence of pesticides in waters ways (yes/no)

1.26 Eating habits. You are preparing to study the eating habits of elementary schoolchildren. Describe two categorical variables and two quantitative variables that you might record for each child. Give the units of measurement for the quantitative variables.

1.27 Mercury in lakes. Mercury is a metal that is highly toxic to the nervous system. Following is a small part of an EESEE data set (“Mercury in Bass”) from a study that assessed the water quality of 53 representative lakes in Florida:[23]

Lake name	pH	Chlorophyll (mg/l)	Avg. mercury in fish (parts per million)	Number of fish sampled	Age of data
Alligator	6.1	0.7	1.23	5	year old
Annie	5.1	3.2	1.33	7	recent
Apopka	9.1	128.3	0.04	6	recent
⋮					

a. What individuals does this data set describe?
b. In addition to the lake’s name, how many variables does the data set contain? Which of these variables are categorical and which are

quantitative?

1.28 Deaths among young people. Here are the number of deaths among persons aged 15 to 24 years in the United States in 2010 due to the leading causes of death for this age group: accidents, 12,015; homicide, 4651; suicide, 4559; cancer, 1594; heart disease, 984; congenital defects, 401.[24]

a. Make two bar graphs of these data: One with bars ordered alphabetically (by death type) and the other with bars ordered from tallest to shortest. Comparisons are easier if you order the bars by height.
b. What additional information do you need to make a pie chart?

1.29 Manatee deaths. Manatees are an endangered species of herbivorous, aquatic mammals found primarily in the rivers and estuaries of Florida. As part of its conservation efforts, the Florida Fish and Wildlife Commission records the cause of death for every recovered manatee carcass. Here is a breakdown of the dead manatee counts in Florida for 2012, by cause of death:[25]

Cause of death	Manatees recovered
Watercraft collisions	81
Perinatal	68
Natural	65
Cold stress	28
Flood gate/canal lock	11
Other human	9
Undetermined	
Total	392

a. Most mortalities recorded as “Undetermined” correspond to manatee carcasses too badly decomposed to make any determination as to the cause of death. How many manatee carcasses had an “undetermined” cause of death?

b. What is the percent of total manatee deaths caused by collisions with watercraft?
c. Make a bar graph sorted by manatee count. What does it show?
d. Could you display these data in a pie chart? Explain why or why not.

1.30 The overweight problem. The 2011 National Health Interview Survey by the National Center for Health Statistics (NCHS) provides weight categorizations for adults 18 years and older based on their body mass index:

Weight category	Percent of adults
Underweight	1.6
Healthy weight	36.3
Overweight	34.2
Obese	27.9

a. Make a bar graph of these data. What do you conclude about the weight problem in the United States?
b. Could you make a single pie chart for these data? (Explain why or why not.) If so, what would it emphasize?

1.31 More on the overweight problem. The same NCHS report (see Exercise 1.30) breaks down the sampled individuals by age group. Here are the percents of obese individuals in the 2011 survey for each age group:

Age group	Percent who are obese
18 to 44	26.2
45 to 64	32.2
65 to 74	31.6
75 and over	19.5

a. Make a bar graph of these data. What do you conclude about the weight problem in the United States?
b. Could you make a single pie chart for these data? (Explain why or why

not.) If so, what would it emphasize?

1.32 Do you have Highly Superior Autobiographical Memory? Some individuals have the ability to recall accurately vast amounts of autobiographical information without mnemonic tricks or extra practice. This ability is called HSAM, for Highly Superior Autobiographical Memory. A research team administered a quiz on past public events to 115 individuals claiming to have HSAM. The histogram in Figure 1.17(a) shows the distribution of scores on this quiz, expressed as percent of correct answers.[26]

a. Describe the distribution of quiz scores for individuals claiming to have HSAM.
b. What does this histogram suggest about the HSAM claims?

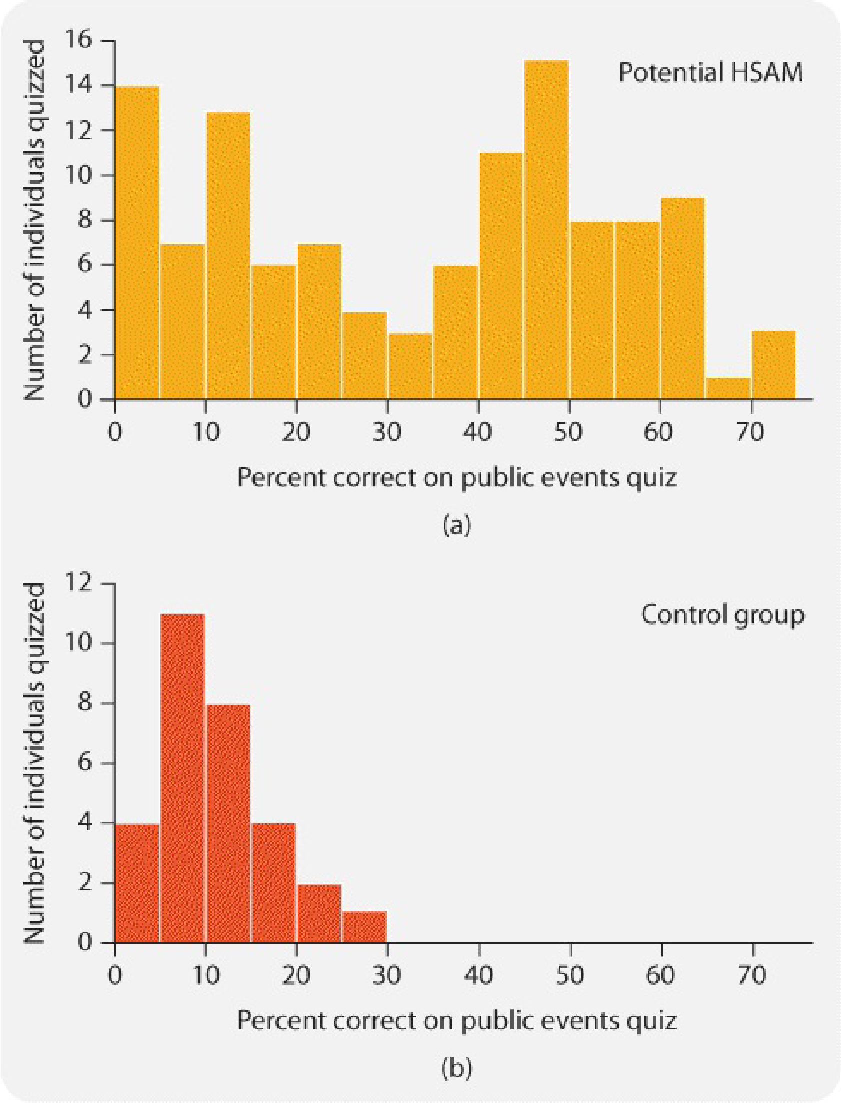


Figure 1.17
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 1.17 Histograms of scores on a quiz of past public events (a) for 115 individuals claiming to have Highly Superior Autobiographical Memory (HSAM) and (b) for 30 control individuals.

1.33 Which graph? Of the following data sets, which would you display with a bar graph and which would you display with a histogram? Explain why.

a. A record of the gender of selected individuals (labeled as male = 0, female = 1)
b. A record of the age of selected parents (in years) at birth of first child
c. A record of the height of selected individuals (in inches)

d. A record of the blood type of selected individuals (A, B, AB, or O)

1.34 Highly Superior Autobiographical Memory, continued. The researchers in Exercise 1.32 also administered the quiz of past public events to 30 individuals who did not claim to have any unusual memory abilities (controls). Their scores are displayed in Figure 1.17(b), using the same horizontal axis scale as that of the histogram of Figure 1.17(a) for easier comparison.

a. Describe the shape, center, and spread of the distribution of quiz scores for the control individuals.
b. Compare the histograms in Figure 1.17(a) and (b). How does the distribution of scores among the control individuals support your interpretation of HSAM claims from Exercise 1.32(b)?

1.35 Acid rain. Changing the choice of classes can change the appearance of a histogram. Here is an example in which a small shift in the classes, with no change in the number of classes, has an important effect on the histogram. The data are the acidity levels (measured by pH) in 105 samples of rainwater. Distilled water has pH 7.00. As the water becomes more acid, the pH goes down. The pH of rainwater is important to environmentalists because of the problem of acid rain.[27]

4.33	4.38	4.48	4.48	4.50	4.55	4.59	4.59	4.61	4.61
4.75	4.76	4.78	4.82	4.82	4.83	4.86	4.93	4.94	4.94
4.94	4.96	4.97	5.00	5.01	5.02	5.05	5.06	5.08	5.09
5.10	5.12	5.13	5.15	5.15	5.15	5.16	5.16	5.16	5.18
5.19	5.23	5.24	5.29	5.32	5.33	5.35	5.37	5.37	5.39
5.41	5.43	5.44	5.46	5.46	5.47	5.50	5.51	5.53	5.55
5.55	5.56	5.61	5.62	5.64	5.65	5.65	5.66	5.67	5.67
5.68	5.69	5.70	5.75	5.75	5.75	5.76	5.76	5.79	5.80
5.81	5.81	5.81	5.81	5.85	5.85	5.90	5.90	6.00	6.03
6.03	6.04	6.04	6.05	6.06	6.07	6.09	6.13	6.21	6.34
6.43	6.61	6.62	6.65	6.81					

a. Make a histogram of pH with 14 classes, using class boundaries 4.2, 4.4, …, 7.0. Describe this histogram. How many peaks does it show?

The presence of more than one peak suggests that the data contain groups that have different distributions.

b. Make a second histogram, also with 14 classes, using class boundaries 4.14, 4.34, …, 6.94. The classes are those from a moved 0.06 to the left. Describe this new histogram. How many peaks does it show?

1.36 Food oils and health. Fatty acids, despite their unpleasant name, are necessary for human health. Two types of essential fatty acids, called omega-3 and omega-6, are not produced by our bodies and so must be obtained from our food. Food oils, which are widely used in food processing and cooking, are major sources of these compounds. Some evidence suggests that a healthy diet should include more omega-3 than omega-6. Table 1.2 gives the ratio of omega-3 to omega-6 in some common food oils.[28] Values greater than 1 show that an oil has more omega-3 than omega-6.

TABLE 1.2 Omega-3 fatty acids as a fraction of omega-6 fatty acids in food oils

Oil	Ratio	Oil	Ratio	Oil	Ratio
Perilla	5.33	Margarine	0.05	Herring	2.67
Walnut	0.20	Olive	0.08	Soybean, hydrogenated	0.07
Wheat germ	0.13	Shea nut	0.06	Rice bran	0.05
Mustard	0.38	Sunflower (oleic)	0.05	Butter	0.64
Sardine	2.16	Sunflower (linoleic)	0.00	Sunflower	0.03
Salmon	2.50	Flaxseed	3.56	Corn	0.01
Mayonnaise	0.06	Canola	0.46	Sesame	0.01
Cod liver	2.00	Soybean	0.13	Cottonseed	0.00
Shortening (household)	0.11	Grape seed	0.00	Palm	0.02
Shortening (industrial)	0.06	Menhaden	1.96	Cocoa butter	0.04

a. Make a histogram of these data, using classes bounded by the whole numbers from 0 to 6.

b. What is the shape of the distribution? How many of the 30 food oils have more omega-3 than omega-6? What does this distribution suggest about the possible health effects of modern food oils?

c. Table 1.2 contains entries for several fish oils (cod, herring, menhaden, salmon, sardine). How do these values support the idea that eating fish is healthy?

1.37 Stemplots: Healing time. Biologists studying the healing of skin wounds measured the rate at which new cells closed a razor cut made to the skin of an anesthetized newt. Here are the sorted data from 18 newts, measured in micrometers per hour:[29]

11	12	14	18	22	22	23	23	26
27	28	29	30	33	34	35	35	40

a. Make a dotplot displaying these data and describe the distribution of healing times.

b. Another type of graphical display used for reasonably small quantitative data sets is the **stemplot.** Stemplots display the information both graphically and numerically by using the numerical values themselves as the basis for the graph. Each data point is split between its final (rightmost) digit and all other digits to its left. These leftmost digits form the *stems* in a vertical column (starting with the smallest at the top) and the last digits are added as *leaves* on a row (left to right, starting with the smallest digit), separated from the stems by a vertical line. Here is how the healing rates appear in a stemplot:

1	1248
2	22336789
3	03455
4	0

Turned 90 degrees counterclockwise, a stemplot looks a lot like a histogram. You can interpret it the same way you would a histogram. Use this stemplot to describe the distribution of healing times.

c. Compare your dotplot with the stemplot shown here, describing their differences and similarities.

1.38 Data maps: Obesity. The CDC studies the obesity crisis in the United States by obtaining the body mass index (BMI, from self-reported weight and height) of representative sets of individuals. A person is considered obese if his or her BMI exceeds 30. Table 1.3 reports the percent of adults in each state who were obese in 2014.[30]

TABLE 1.3 Percent of the population who is obese, for each state and the Distr Columbia

State	Percent	State	Percent	State	Percent	State
Alabama	33.5	Illinois	29.3	Montana	26.4	Rhode Island
Alaska	29.7	Indiana	32.7	Nebraska	30.2	South Carolina
Arizona	28.9	Iowa	30.9	Nevada	27.7	South Dakota
Arkansas	35.9	Kansas	31.3	New Hampshire	27.4	Tennesse
California	24.7	Kentucky	31.6	New Jersey	26.9	Texas
Colorado	21.3	Louisiana	34.9	New Mexico	28.4	Utah
Connecticut	26.3	Maine	28.2	New York	27.0	Vermont
Delaware	30.7	Maryland	29.6	North Carolina	29.7	Virginia
D.C.	21.7	Massachusetts	23.3	North Dakota	32.2	Washingt
Florida	26.2	Michigan	30.7	Ohio	32.6	West Virginia
Georgia	30.5	Minnesota	27.6	Oklahoma	33.0	Wisconsi
Hawaii	22.1	Mississippi	35.5	Oregon	27.9	Wyomin
Idaho	28.9	Missouri	30.2	Pennsylvania	30.2	

a. Define the individuals and the variable shown in Table 1.3. Is the variable quantitative or categorical?

b. Make a dotplot or a histogram of the obesity data and describe the shape, center, and spread of the distribution. What do the data reveal about obesity in the United States?

c. Just as data collected over time should be displayed on a time plot to look for evidence of patterns over time, so data collected geographically should be plotted on a map to look for evidence of broad geographical patterns. Figure 1.18 displays the state data on a map of the United States. What geographical pattern can you identify from this map? Would you have been able to identify this parttern from your previous graph or from the data table alone?

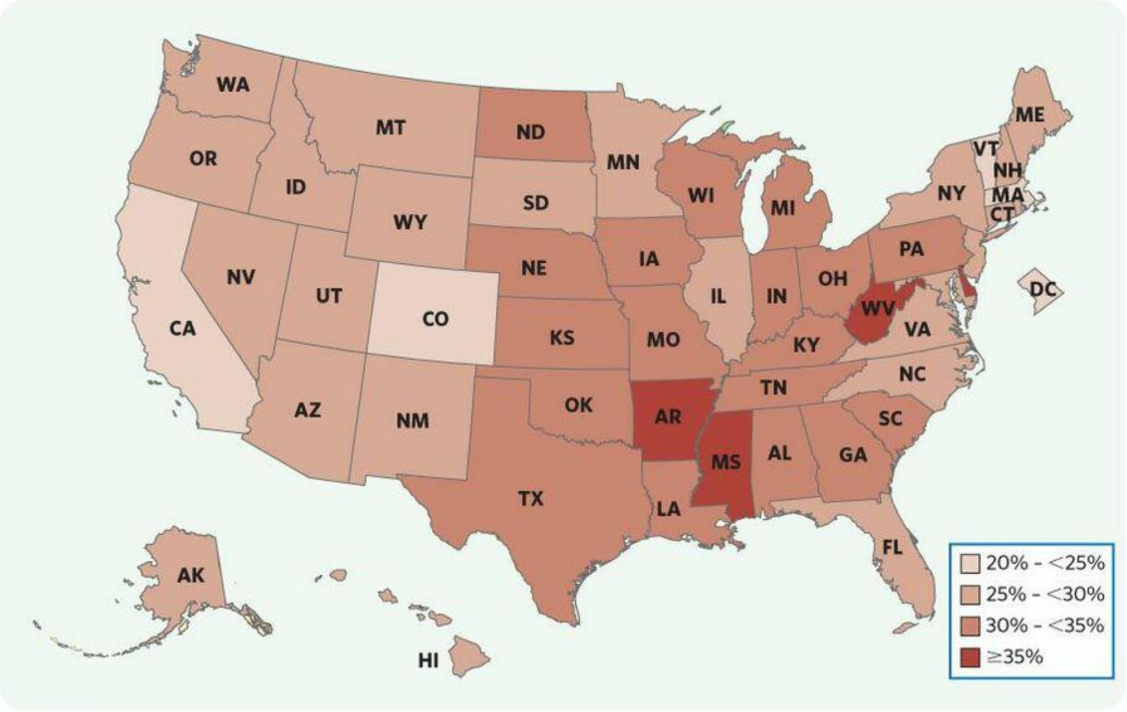

WA
OR
CA
NV
ID
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
KY
TN
MS
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
ME
VT
NH
MA
CT
DC
AK
HI
20% - <25%
25% - <30%
30% - <35%
≥35%

Figure 1.18
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.18 Percent of the adult population who is obese, plotted for each state on a map of the United States

1.39 Stemplots: Flower length. Here are the flower lengths (in millimeters) of 16 specimens of the tropical plant *Heliconia bihai.*[31]

46.3	46.4	46.6	46.7	46.8	46.8	46.9	47.1
47.1	47.4	48.1	48.2	48.3	48.4	50.1	50.3

a. Follow the model in Exercise 1.37 to create a stemplot of these flower lengths.
b. Use your stemplot to describe the distribution of flower lengths in this tropical plant.

1.40 Fur seals on Saint George Island. Every year hundreds of thousands of northern fur seals return to their haul-out territory in the Pribilof Islands in Alaska to breed, give birth, and teach their pups to swim, hunt, and survive in the Bering Sea. U.S. commercial fur sealing operations ended in 1983, but despite the reduction in harvest, the population of fur seals has continued to decline. Here are data on the number of fur seal pups born on Saint George Island (in thousands) from 1975 to 2006:[32]

Year	Pups born (thousands)	Year	Pups born (thousands)
1975	53.70	1991	24.28
1976	56.16	1992	25.16
1977	43.41	1993	23.70
1978	47.25	1994	22.24
1979	47.47	1995	24.82
1980	39.34	1996	27.39
1981	38.15	1997	24.74
1982	39.29	1998	22.09
1983	31.44	1999	21.13
1984	33.44	2000	20.18
1985	28.87	2001	18.89
1986	32.36	2002	17.59
1987	33.12	2003	17.24
1988	24.82	2004	16.88
1989	33.11	2005	16.97
1990	23.40	2006	17.07

Make a dotplot or a histogram to display the distribution of pups born per year. Describe the shape, center, and spread of the distribution. Are there any outliers?

1.41 Nanomedicine. Researchers examined a new treatment for advanced ovarian cancer in a mouse model. They created a nanoparticle-based delivery system for a suicide gene therapy to be delivered directly to the tumor cells. The grafted tumors were injected either with the new treatment or with only some buffer solution to serve as a comparison. The following data give the fold increase in tumor size after two weeks in 20 mice. A 1 represents no change; a 2 represents a doubling in volume of the tumor.[33]

Buffer solution									
9.1	8.1	7.8	7.0	6.8	5.4	5.4	4.1	3.8	3.3
Nanoparticle-delivered gene therapy									
4.1	3.5	2.1	2.1	1.8	1.8	1.4	1.2	1.1	1.1

a. Make two dotplots, one for each group, using the same scale on the horizontal axis for both. Describe the distribution of tumor increase in each treatment group.

b. Report the approximate midpoints of both groups. What are the most important differences between the two groups? What can you conclude from the study findings?

1.42 Fur seals on Saint George Island, continued. Make a time plot of the number of fur seals born per year from Exercise 1.40. What does the time plot show that your plot in Exercise 1.40 does not show? When you have data collected over time, a time plot is often needed to understand what is happening.

1.43 Herbicide resistance in weeds. Farmers use herbicides to limit the growth of weeds among their crops. Eventually, this technique places evolutionary pressure on weed species, resulting in herbicide resistance. The International Survey of Herbicide Resistant Weeds keeps track of documented cases of herbicide resistance in 80 countries worldwide. Each case is examined to identify the weed species and particular mutation allowing herbicide resistance. Table 1.4 gives the number of unique cases of herbicide resistance (corresponding to a specific weed species and a specific mutation) documented worldwide every year since 1950.[34]

TABLE 1.4 **Number of unique cases of herbicide resistance in weeds worldwide, by year**

Year	Number of unique cases	Year	Number of unique cases	Year	Number of unique cases	Year	Number of unique cases	Year	Number of unique cases
1950	0	1963	2	1976	12	1989	115	2002	289

1951	0	1964	3	1977	14	1990	129	2003	293
1952	0	1965	3	1978	18	1991	139	2004	311
1953	0	1966	3	1979	27	1992	151	2005	331
1954	0	1967	3	1980	41	1993	170	2006	335
1955	0	1968	3	1981	49	1994	184	2007	344
1956	0	1969	3	1982	60	1995	191	2008	350
1957	2	1970	4	1983	65	1996	211	2009	365
1958	2	1971	4	1984	72	1997	225	2010	377
1959	2	1972	5	1985	76	1998	239	2011	391
1960	2	1973	7	1986	80	1999	256	2012	397
1961	2	1974	7	1987	92	2000	269		
1962	2	1975	9	1988	99	2001	278		

a. Make a time plot of the data. Does the time plot illustrate only year-to-year variations or are there other patterns apparent? Specifically, is there a trend over any period of years? What about cyclical fluctuation? Explain in words the change in the worldwide number of unique cases of herbicide resistance in weeds over this 63-year period.

b. Do you think it would be appropriate to display these data in a histogram without first plotting them in a time plot? Explain why or why not.

1.44 Exercising. The Gallup polling organization interviews samples of individuals to learn about public opinions and habits. One question Gallup asks regularly of randomly picked Americans is whether they exercised for at least 30 minutes on three or more of the past seven days. Figure 1.19 is a time plot of the percent of respondents who say that they did, for surveys conducted between 2008 and 2015.[35]

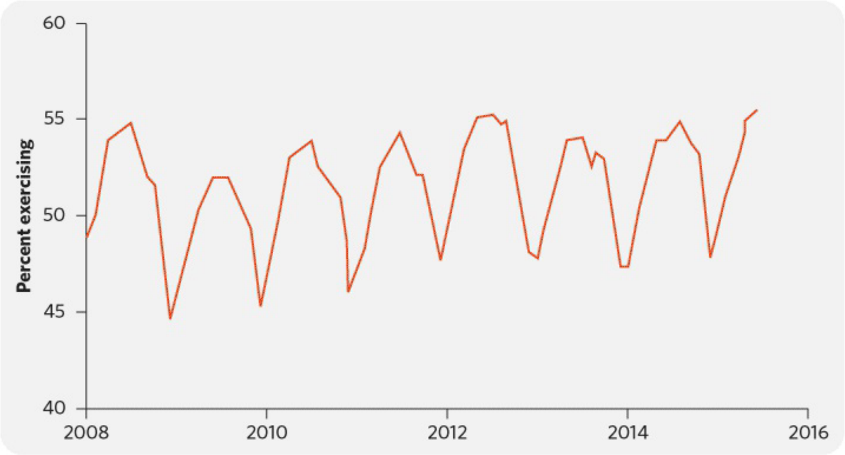


Figure 1.19
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 1.19 Time plot of the percent of respondents in numerous Gallup polls who say that they exercised at least 30 minutes on three or more days in the previous week.

a. The most notable pattern in this time plot is yearly cycles. In which season of the year is the percent who exercise regularly highest? Lowest? (To read the graph, note that the tick mark for each year is at the beginning of the year.)

b. Is there a longer-term trend visible in addition to the cycles? If so, describe it.

Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Short answers are not available for these exercises to enable their use for class work or assignments.*

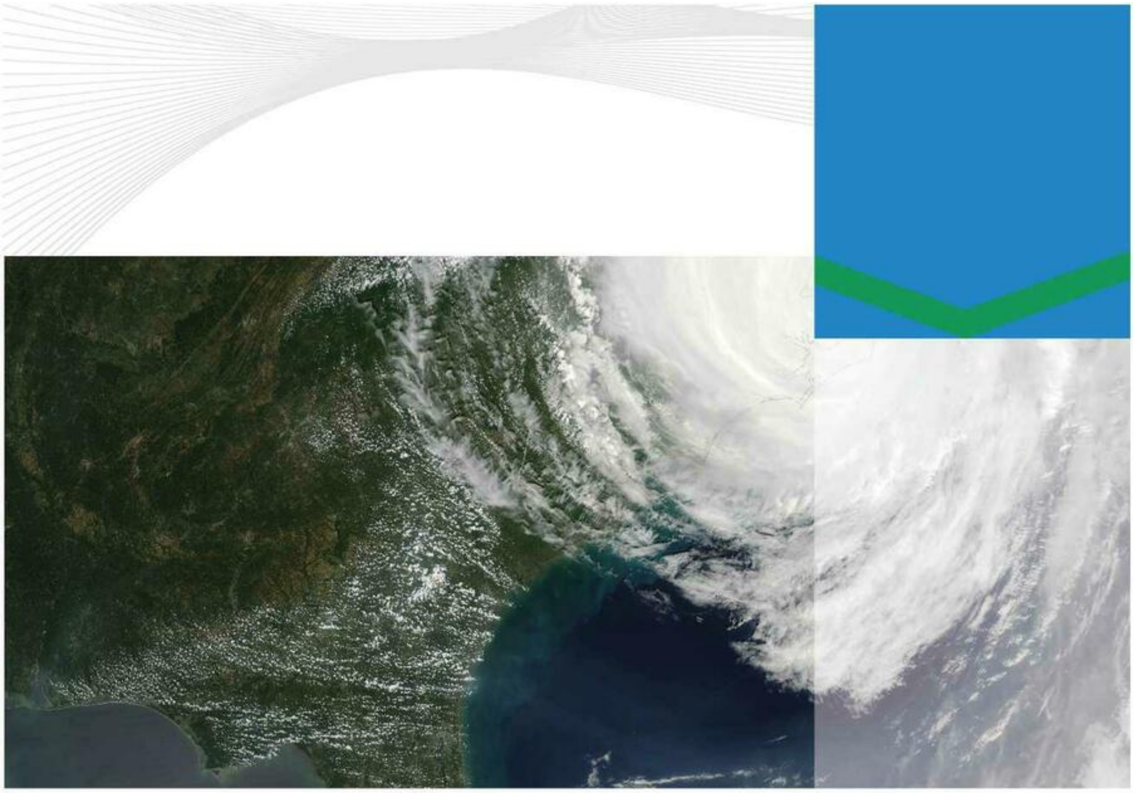
1.45 Everglades. Everglades National Park is the largest subtropical wilderness in the United States and has been designated a Wetland of International Importance. This important ecosystem is closely monitored by the U.S. Geological Survey. The *Large.Everglades* data file contains data on five variables describing the properties of the Shark River at the Gunboat Island station (on the coast of the Gulf of Mexico) over a seven-day period ranging from November 25 to December 2, 2009. The data are recordings taken every 15 minutes.[36]

a. Make a histogram for each of the five variables and describe each distribution.

b. Make a time plot for each of the five variables and describe any pattern in the data.

c. Summarize your findings and describe some of the key features of the river at this station by the ocean.

CHAPTER 2 Describing Quantitative Distributions with Numbers



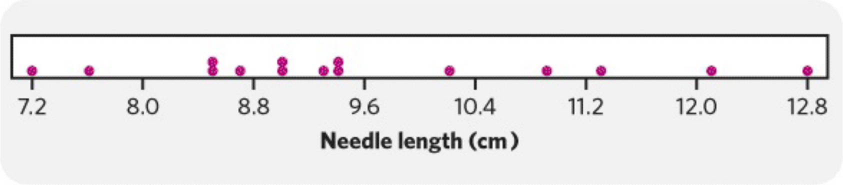
NASA

IN THIS CHAPTER WE COVER...

- Measures of center: median, mean
- Measures of spread: percentiles, standard deviation
- Graphical displays of numerical summaries
- Spotting suspected outliers
- *Discussion: Dealing with outliers*
- Organizing a statistical problem

In Chapter 1 we discussed how the techniques of exploratory data analysis apply equally to data sets obtained from a given *population* (the entire group of individuals about which we want information) or from only those individuals in a smaller *sample*. The distinction between population and sample is an important one when making statistical inferences, which is why different terms and notations are typically used to distinguish them. Numerical summaries are called **parameters** when they describe an entire population and **statistics** when they describe only a sample. In this chapter, we introduce the specific notations and computations for summary statistics relating to samples. We will discuss parameters in more detail starting in Part II of this book.

Examining nature closely reveals surprising variability. The needles in a given pine tree species are not all the same size, for instance. It is the distribution of needle lengths, then, that characterizes the pine species. The following is a dotplot of the lengths (in centimeters, cm) of a sample of 15 needles taken at random from different parts of several Aleppo pine trees located in Southern California:[1]



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

The distribution is single-peaked and slightly skewed to the right without outliers. Our goal in this chapter is to use numbers to describe the center and the spread of this and other distributions of *quantitative variables*. Categorical variables are simply summarized by the count or proportion (percent) of the data falling into each category, as we saw in Chapter 1.

MEASURES OF CENTER: MEDIAN, MEAN

In Chapter 1, we used the approximate midpoint of a distribution (displayed in a dotplot or a histogram) as an informal measure of center. The *median* is the formal version of the midpoint, with a specific rule for calculation. It is the simplest measure of center.

THE MEDIAN *M*

The **median M**M is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.
2. If the number of observations nn is odd, the median MM is the center observation in the ordered list. Find the location of the median by counting (n+1)/2$(n+1)/2$ observations up from the smallest observation in the list.
3. If the number of observations nn is even, the median MM is the mean of the two center observations in the ordered list. The location of the median is again (n+1)/2,$(n+1)/2,$ counting from the smallest observation in the list.

Note that the formula (n+1)/2$(n+1)/2$ does *not* give the median, but rather the *location* of the median in the ordered list. Finding the median requires little arithmetic, so it is easy to do by hand for small sets of data. You can even find the median when the data are summarized in a frequency table, as illustrated in Exercise 2.3. In practice, most people rely on a calculator or statistical software to obtain the median and other numerical summaries.

EXAMPLE 2.1 Finding the median: odd *n*

Craig Tuttle/Corbis Documentary/Getty Images

What is the median length for our 15 Aleppo pine needles? Here are the data, arranged in increasing order:

7.2 7.6 8.5 8.5 8.7 9.0 9.0 **9.3** 9.4 9.4
10.2 10.9 11.3 12.1 12.8

The number of observations $n = 15$ is odd. The bold **9.3** is the center observation in the ordered list, with 7 observations to its left and 7 observations to its right. This value is the median, $M = 9.3\text{ cm}$.

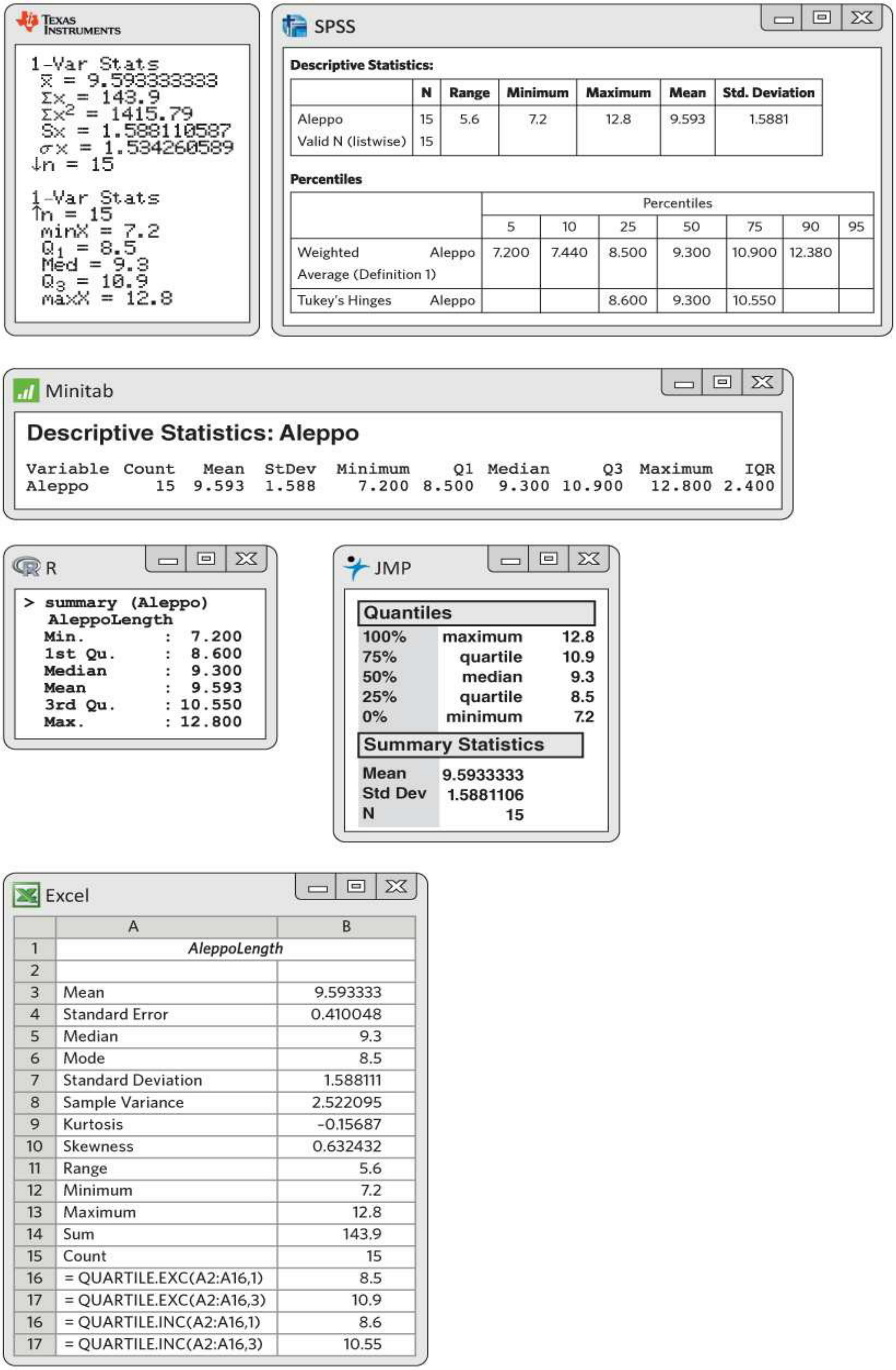
Because $n = 15$, our rule for the location of the median gives

location of M=n+12=162=8

$$\text{location of } M = \frac{n+1}{2} = \frac{16}{2} = 8$$

That is, the median is the 8th observation in the ordered list. Use this rule to locate the center in an ordered list or even in a dotplot.

Figure 2.1 displays output describing the 15 Aleppo pine needle lengths using a graphing calculator, a spreadsheet program, and four statistical software packages. Most refer to the median as “Median” (or “Med”), although SPSS labels it “Percentile 50.” The median is, in fact, the 50th **percentile** (JMP calls it a “quantile”) of a data set because it is the value such that half (50%) of the observations are smaller and the other half are larger. Differences in output from different statistical programs can be frustrating at first, but they are typically superficial. Once you know what to look for, you will be able to correctly interpret output from any technological tool.



```
1-Var Stats
 x̄ = 9.593333333
 Σx = 143.9
 Σx² = 1415.79
 Sx = 1.588110587
 σx = 1.534260589
↓n = 15

1-Var Stats
↑n = 15
 minX = 7.2
 Q₁ = 8.5
 Med = 9.3
 Q₃ = 10.9
 maxX = 12.8
```

Descriptive Statistics:

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Aleppo	15	5.6	7.2	12.8	9.593	1.5881
Valid N (listwise)	15					

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Aleppo	7.200	7.440	8.500	9.300	10.900	12.380	
Tukey's Hinges	Aleppo			8.600	9.300	10.550		

Descriptive Statistics: Aleppo

```
Variable Count  Mean  StDev  Minimum    Q1  Median     Q3  Maximum   IQR
Aleppo      15 9.593  1.588    7.200 8.500   9.300 10.900   12.800 2.400
```

```
> summary (Aleppo)
  AleppoLength
 Min.      :  7.200
 1st Qu.   :  8.600
 Median    :  9.300
 Mean      :  9.593
 3rd Qu.   : 10.550
 Max.      : 12.800
```

```
Quantiles
100%   maximum    12.8
75%    quartile   10.9
50%    median      9.3
25%    quartile    8.5
0%     minimum     7.2
Summary Statistics
Mean     9.5933333
Std Dev  1.5881106
N               15
```

	A	B
1	*AleppoLength*	
2		
3	Mean	9.593333
4	Standard Error	0.410048
5	Median	9.3
6	Mode	8.5
7	Standard Deviation	1.588111
8	Sample Variance	2.522095
9	Kurtosis	-0.15687
10	Skewness	0.632432
11	Range	5.6
12	Minimum	7.2
13	Maximum	12.8
14	Sum	143.9
15	Count	15
16	= QUARTILE.EXC(A2:A16,1)	8.5
17	= QUARTILE.EXC(A2:A16,3)	10.9
16	= QUARTILE.INC(A2:A16,1)	8.6
17	= QUARTILE.INC(A2:A16,3)	10.55

Figure 2.1

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.1 Output from a graphing calculator, a spreadsheet program, and four software packages describing the Aleppo pine needle length data.

EXAMPLE 2.2 Finding the median: even nn

We also have the lengths (in cm) of 18 needles from trees of the noticeably different Torrey pine species. What is the median length for these 18 pine needles? The ordered data are:

21.2 21.6 21.7 23.1 23.7 24.2 24.2 25.5 **26.6** **26.8**
28.9 29.0 29.7 29.7 30.2 32.5 33.7 33.7

There is no unique center observation, but there is a center pair—the bold values **26.6** and **26.8**, which have 8 observations before them in the ordered list and 8 observations after them. The median is midway between these two observations:

M=26.6+26.82=26.7 cm

$$M = \frac{26.6+26.8}{2} = 26.7\,\text{cm}$$

With n=18,$n = 18,$ the rule for locating the median in the list gives

location of M=n+12=192=9.5

$$\text{location of } M = \frac{n+1}{2} = \frac{19}{2} = 9.5$$

The location 9.5 means “halfway between the 9th and 10th observations in the ordered list.”

Another important measure of center is the ordinary arithmetic average, or *mean*. The mean is the most commonly reported measure of center.

THE MEAN x-$\bar{x}$

To find the **mean** of a set of observations, add their values and divide by the number of observations. If the nn observations are x1,x2,… ,xn,$x_1, x_2, \dots, x_n,$ their mean is

x¯=x1+x2+⋯+xnn

$$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n}$$

or, in more compact notation,

$$\bar{x}=\frac{1}{n}\sum x_i$$

$$\bar{x} = \frac{1}{n}\sum x_i$$

The ∑ (capital Greek letter sigma) in the formula for the mean is the math symbol for summation. The subscripts on the observations x_i are just a way of keeping the n observations distinct. They do not necessarily indicate order or any other special facts about the data. The bar over the x indicates the mean of all the x-values; the mean $\bar{x}$ is pronounced "x-bar." This notation is very common. When writers who are discussing data use $\bar{x}$ or $\bar{y}$, they are talking about a mean.

EXAMPLE 2.3 Mean of pine needle lengths

The mean needle length of our 15 Aleppo pine needles is

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{7.2+7.6+\cdots+12.8}{15}=\frac{143.9}{15}=9.59 \text{ cm}$$

$$\begin{aligned}\bar{x} &= \frac{x_1+x_2+\cdots+x_n}{n}\\ &= \frac{7.2+7.6+\cdots+12.8}{15}=\frac{143.9}{15}=9.59\text{ cm}\end{aligned}$$

In practice, to obtain the mean for a data set, you will key the data into your calculator or software and ask for the mean alone or in a set of summary statistics. Figure 2.1 shows the mean labeled as either "Mean" or "$\bar{x}$," depending on the technology.

Paying attention to the formula and step-by-step computations is nonetheless useful. It allows us to see that, unlike the median, the mean uses *all* the values in the data set. Therefore, very small or very large numerical values, such as those found in skews and outliers, will influence the value of the mean. Because the mean cannot resist the influence of extreme observations, we say that it is *not* a **resistant measure** of center. The median, in contrast, is a resistant measure of center because it is influenced only by the total number of data points and the numerical value of the point or points located at the center of the distribution. The *Mean and Median* applet is an excellent way to compare the resistance of MM and x¯ .$\bar{x}$.

COMPARING THE MEAN AND THE MEDIAN

The mean and the median of a symmetric distribution are close together. If the distribution is exactly symmetric, the mean and the median are exactly the same. In a skewed distribution, the mean is usually farther out in the long tail than is the median.[2]

EXAMPLE 2.4 Flooding in coastal communities

Figure 2.2 shows a dotplot of the number of flood events recorded between 2009 and 2013 in each of 52 Atlantic coastal communities between Maine and Texas.[3] Flood events were defined as days when recorded tide heights exceeded local water level thresholds, triggering flood conditions. The distribution is noticeably skewed to the right and has some potential high outliers. The mean number of flood events for this time period was 55.3, whereas the median was only 37.5. The mean is pulled toward the right tail of this right-skewed distribution.

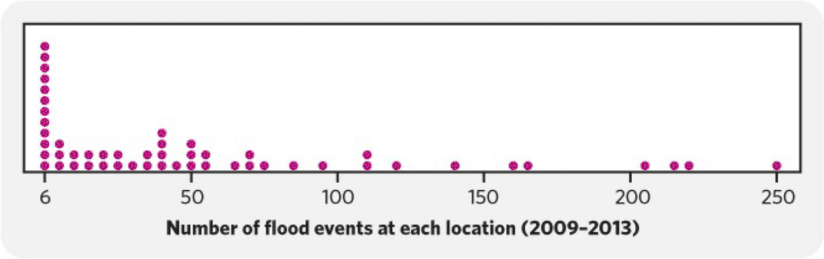


Figure 2.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.2 Dotplot of the number of flood events recorded over the 2009-2013 period for each of 52 coastal communities.

NASA

If the largest number of flood events were increased tenfold to 2500 events rather than 250 events, the mean would increase to 98.6 but the median would not change at all. The artificially created outlier would count as just one observation above the center, no matter how far above the center it might lie. In contrast, the mean uses the actual value of each observation and, therefore, is very sensitive to any extreme values.

Many biological variables have distributions that are skewed to the right. Survival times, for instance, are typically right-skewed. Most individuals die within a certain period of time, although a few will survive much longer. When dealing with strongly skewed distributions, it is somewhat customary to report the median (“midpoint”) rather than the mean (“arithmetic average”) because the median gives an idea of the center of the “typical” part of the distribution. However, a health organization or a government agency may need to account for all possible survival times, and therefore calculate the mean, to estimate the cost of medical care for a given disease and to plan medical staffing appropriately. Relying only on the median would result in underestimating the medical and financial needs. The mean and median measure center in different ways, and both are useful.

APPLY YOUR KNOWLEDGE

2.1 Spider silk. Spider silk is the strongest known material, natural or man-made, on a weight basis. A study examined the mechanical properties of spider silk using 21 female golden orb weavers, *Nephila* *clavipes*. Here are data on silk yield stress, which represents the amount of force per unit area needed to reach permanent deformation of the silk strand. The data are expressed in megapascals (MPa).[4]

164.0	478.7	251.3	351.7	173.0	448.9	300.6
362.0	272.4	740.2	329.0	327.2	270.5	332.1
288.8	176.1	282.2	236.1	358.2	270.5	290.7

a. Make a dotplot of these data. Describe the shape, center, and spread of the distribution.
b. Find the mean and median yield stress. Compare these two values. Referring to your plot, what general fact does your comparison illustrate?

2.2 Deep-sea sediments. Phytopigments are markers of the amount of organic matter that settles in sediments on the ocean floor. Phytopigment concentrations in deep-sea sediments collected worldwide showed a very strong right-skew.[5] Of two summary statistics, 0.015 and 0.009 gram per square meter of bottom surface, which one is the mean and which one is the median? Explain your reasoning.

2.3 Age of participants, summarized. A study of a new type of vision screening test recruited a sample of 175 children age 3 to 7 years. The publication provides the following summary of the children's ages: "Twelve patients (7%) were 3 years old; 33 (19%), 4 years old; 29 (17%), 5 years old; 69 (39%), 6 years old; and 32 (18%), 7 years old."[6]

a. What is the median age in the study? Notice that you can easily add up the percents provided in parentheses (in increasing order of age) until the total just exceeds 50%.
b. What is the mean age in the study? You will need to either organize the data in a way that your technology will accept or do the computations by hand. If so, be sure to multiply each age by the number of children with that age in the numerator of the formula for the mean.
c. Display the reported ages in a histogram. Compare the values of

the mean and the median in relation to the age distribution in the study. What general fact does your comparison illustrate?

MEASURES OF SPREAD: PERCENTILES, STANDARD DEVIATION

In Chapter 1, we described the distribution of a quantitative variable in a histogram or a dotplot by both its center and its spread. Similarly, *the numerical summaries we provide for a quantitative variable should offer both a measure of center and a measure of spread.*

One way to measure spread is to give the smallest and largest observations in the distribution. For example, the number of coastal flood events in Example 2.4 ranged from 0 to 250. The distance between the minimum and the maximum values in a distribution is called the **range;** in this case, it is 250 (250 − 0).

The minimum and the maximum show the full spread of the data, but they may actually be outliers. Another, more resistant way to describe the spread of a quantitative variable is to look at the spread of the middle half of the data. The first and third *quartiles* mark out the central half of the distribution. The *interquartile range* is the distance between the first and third quartiles.

When the list of observations is sorted in increasing order, the *first quartile* lies one-quarter of the way up the list, and the *third quartile* lies three-quarters of the way up the list. In other words, the first quartile is larger than 25% of the observations, and the third quartile is larger than 75% of the observations. Note that the 50th percentile (the median) is the second quartile, which is larger than 50% of the observations. That is the key idea for quartiles and any other percentiles.

You would typically use technology to obtain summary statistics, but here is how to find the quartiles of a data set by hand.

THE QUARTILES Q1Q_1 AND Q3Q_3

To calculate the **quartiles:**

1. Arrange the observations in increasing order and locate the median MM in the ordered list of observations.
2. The **first quartile Q1**Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the

overall median.

3. The **third quartile Q3**Q_3 is the median of the observations whose position in the ordered list is to the right of the location of the overall median.

The **interquartile range *IQR*** is the distance between the first and third quartiles:

IQR=Q3-Q1

$$IQR = Q_3 - Q_1$$

The next two examples show how the rules for the quartiles work for both odd and even numbers of observations.

EXAMPLE 2.5 Finding the quartiles: odd *n*

Our sample of 15 Aleppo pine needles (Example 2.1), arranged in increasing order, is:

7.2 7.6 8.5 8.5 8.7 9.0 9.0 **9.3** 9.4 9.4
10.2 10.9 11.3 12.1 12.8

We have an odd number of observations, so the median is the middle one, the bold **9.3** in the list. The first quartile is the median of the 7 observations to the left of the median. This is the 4th of these 7 observations, so Q1=8.5 cm.$Q_1 = 8.5\text{ cm}$. If you want, you can use the formulation for locating the median with n=7:$n = 7$:

location of Q1=n+12=7+12=4

$$\text{location of } Q_1 = \frac{n+1}{2} = \frac{7+1}{2} = 4$$

The third quartile is the median of the 7 observations to the right of the median, Q3=10.9 cm.$Q_3 = 10.9\text{ cm}$.

The quartiles are resistant to outliers. For example, Q3Q_3 would still be 10.9 if the largest needle length were 50 cm rather than 12.8 cm.

Look at the output displays for these 15 Aleppo pine needle lengths in Figure 2.1. The TI-83, JMP, and Minitab all agree with our work. The R output, however, says that Q1=8.6$Q_1 = 8.6$ (“1st Qu.”) and Q3=10.55$Q_3 = 10.55$ (“3rd Qu.”). How can this be? *There are several rules for finding the quartiles. SPSS provides the results from both methods (“Percentiles 25 and 75”). Some software packages use rules that give results different from ours for some sets of data.* Our rule is the simplest way to perform the computations by hand. Results from applying the various rules are always close to each other, however, so *to describe data you should just use the answer your technology gives you.* Excel’s “Descriptive Statistics” menu item gives more information than we typically need (such as the “mode,” the most common numerical value in the data set), but it does not provide the quartiles. They can be obtained from a separate quartile function, which offers a choice of two computation approaches (as does SPSS).

EXAMPLE 2.6 Finding the quartiles: even *n*

Here are the lengths of the 18 Torrey pine needles (Example 2.3), arranged in increasing order:

21.2 21.6 21.7 23.1 23.7 24.2 24.2 25.5 26.6 * 26.8
28.9 29.0 29.7 29.7 30.2 32.5 33.7 33.7

We have an even number of observations, so the median lies midway between the middle pair, the 9th and 10th values in the list. The median value is M=26.7 cm$M = 26.7\ \text{cm}$ and its location is marked by a star. The first quartile is the median of the first 9 observations, because these are the observations to the left of the location of the median. Confirm that Q1=23.7 cm$Q_1 = 23.7\ \text{cm}$ and Q3=29.7 cm.$Q_3 = 29.7\ \text{cm}.$

Be careful when several observations take the same numerical value, as in these examples and in Exercise 2.3. Include all the observations and apply the rules just as if they all had distinct values.

The minimum, first quartile, median, third quartile, and maximum are commonly reported together as the *five-number summary*. These five numbers offer a reasonably complete description of a data set’s center and spread.

THE FIVE-NUMBER SUMMARY

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

$$\text{Minimum} \quad Q_1 \quad M \quad Q_3 \quad \text{Maximum}$$

The five-number summaries from Examples 2.5 and 2.6 are

Aleppo pine	7.2	8.5	9.3	10.9	12.8
Torrey pine	21.2	23.7	26.7	29.7	33.7

The five-number summary is not the most common numerical description of a distribution. Rather, that distinction belongs to the combination of the mean (a measure of center) and the *standard deviation* (a measure of spread). The standard deviation and its square value, the *variance*, measure spread by looking at how far the observations are from their mean.

THE SAMPLE STANDARD DEVIATION s

The **variance** s^2 of a *sample* set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of n sample observations $x_1, x_2, \ldots, x_n$ is

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}$$

or, more compactly,

$$s^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$$

$$s^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2$$

The **sample standard deviation s**s is the square root of the sample variance s2:s^2 :

s=1n-1∑(xi-x¯)2

$$s=\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}$$

In practice, you will always use software or your calculator to obtain the standard deviation from keyed-in data. However, doing the calculations step by step once helps understand how the variance and standard deviation measure spread.

EXAMPLE 2.7 Calculating the standard deviation

PJF Military Collection/Alamy Stock Photo

A person’s metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. The units are kilocalories (Cal) for a 24-hour period, where kilocalories are the same calories used to describe the energy content of foods.

1792 1666 1362 1614 1460 1867 1439

The researchers reported x¯=1600 Cal$\bar{x}=1600\text{ Cal}$ and s=189.2 Cal $s=189.2\text{ Cal}$ for this sample of 7 men. Here is how they were computed, step by step. First find the mean:

x¯=1792+1666+1362+1614+1460+1867+14397=11,2007=1600 Cal

$$\begin{aligned}\bar{x} &= \frac{1792+1666+1362+1614+1460+1867+1439}{7}\\ &= \frac{11{,}200}{7} = 1600\ \text{Cal}\end{aligned}$$

Figure 2.3 displays the data as points above the number line, with the sample mean x¯ $\bar{x}$ marked by an asterisk (*). The arrows mark two of the deviations from the mean. These deviations show how spread out the data are about their mean. They are the starting point for calculating the variance and the standard deviation.

Observations **xi**x_i	**Deviations** **xi−x¯**$x_i-\bar{x}$			**Squared deviations** **(xi−x¯)2**$(x_i-\bar{x})^2$		
1792	1792−1600	=	192	192^2	=	36,864
1666	1666−1600	=	66	66^2	=	4,356
1362	1362−1600	=	−238	$(-238)^2$	=	56,644
1614	1614−1600	=	14	14^2	=	196
1460	1460−1600	=	−140	$(-140)^2$	=	19,600
1867	1867−1600	=	267	267^2	=	71,289
1439	1439−1600	=	−161	$(-161)^2$	=	25,921
	sum	=	0	sum	=	214,870

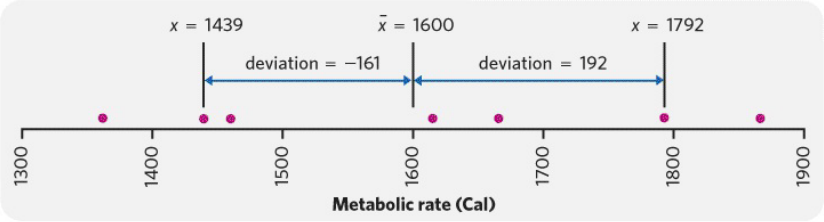


Figure 2.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.3 Metabolic rates for 7 men, with their mean (*) and the deviations of two observations from the mean indicated.

The sample variance is the sum of the squared deviations divided by 1 less than the number of observations:

s2=214,8706=35,811.67 Cal2

$$s^2 = \frac{214{,}870}{6} = 35{,}811.67\ \text{Cal}^2$$

The sample standard deviation is the square root of the sample variance:

> $$s = \sqrt{35,811.67} = 189.24 \text{ Cal}$$

Notice that the “average square deviation” in the sample variance s^2 actually divides the sum by 1 less than the number of observations—that is, by $n - 1$ rather than by n. The reason is that the deviations $x_i - \overline{x}$ always sum to exactly 0, so knowing $n - 1$ of them determines the last one. Only $n - 1$ of the squared deviations in a sample can vary freely once we know $\overline{x}$, and we find their average by dividing the total by $n - 1$. This adjustment is necessary to give the sample variance important mathematical properties that will be addressed in later chapters on statistical inference. The number $n - 1$ is called the **degrees of freedom** of the sample variance or sample standard deviation. Some calculators and software offer a choice between dividing by n and dividing by $n - 1$, so be sure to use $n - 1$ when calculating the sample standard deviation. For instance, the TI-83 calculator output in Figure 2.1 gives both the sample standard deviation, “$\text{S}_\text{X} = 1.588...$,” and the value you would get using n in the denominator, “$\sigma\text{x} = 1.534...$” (We will see in Chapter 13 that the notation for the population standard deviation is σ.)

More important than the details of hand calculation are the properties that determine the usefulness of the sample standard deviation:

- s measures *spread about the mean*[7] and should be used only when the mean is chosen as the measure of center. Although this is a gross simplification, you might find it helpful to think of the standard deviations as representing roughly the average dispersion, in either direction, of the n data points relative to their mean.

- s is *always zero or greater than zero.* $s = 0$ only when there is no spread (the values in the sample are all identical). Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.

- s has the *same units of measurement as the original observations*. For example, if you measure metabolic rates in kilocalories, both the mean $\overline{x}$ and the standard deviation s are also in kilocalories. This is one reason to prefer ss to the variance s2,s^2, which in this case is measured in squared kilocalories (cal2).(cal^2). The variance, however, has more interesting mathematical properties that are beyond the scope of this chapter.

- Like the mean x¯,$\bar{x}$, ss is *not resistant*. Outliers and skew increase the spread of a distribution and, therefore, also increase s.s.

CAUTION

The use of squared deviations renders s even more sensitive than x¯$\bar{x}$ *to a few extreme observations.* In Example 2.4, we saw that the mean number of flood events between 2009 and 2013 for the 52 East Coast communities was x¯=52.3,$\bar{x} = 52.3$, but that changing the largest observation from 250 to 2500 would increase the mean to 98.6. This hypothetical typo would also result in a much larger standard deviation of 344.6 instead of the true standard deviation s=65.0.$s = 65.0$.

APPLY YOUR KNOWLEDGE

2.4 Mercury levels in pregnant women. A study recorded the blood mercury levels (in micrograms per liter, μg/l) of 4134 pregnant British women enrolled in the Avon Longitudinal Study of Parents and Children.[8] The published findings include the following statement: "Blood mercury levels ranged from 0.17 to 12.8 μg/l. The 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles were 0.81, 0.99, 1.35, 1.86, 2.52, 3.33, and 4.02 μg/l, respectively."

a. Find the five-number summary of the distribution of blood mercury levels among the pregnant women enrolled in this study.
b. What are the range and the interquartile range for this distribution?
c. Do you have enough information to compute the standard deviation of this sample of blood mercury levels? Explain your reasoning.

2.5 Spider silk, continued. In Exercise 2.1 you plotted the silk yield stress for 21 female golden orb weaver spiders.

a. Obtain the five-number summary of the distribution of yield stresses.

b. Obtain the mean and the standard deviation of the sample of yield stresses.
c. Which summary gives more information about the distribution of silk yield stresses? How do they reflect what you see in your dotplot? Remember that a summary for a quantitative variable, no matter how detailed, will never be as informative as a graph of the raw data.

Jonathan Heger/istockphoto

GRAPHICAL DISPLAYS OF NUMERICAL SUMMARIES

Numerical values are always more challenging to interpret and communicate than graphical displays. For this reason, we also like to display numerical summaries in graphical format.

The five-number summary of a distribution can be displayed in a *boxplot*. Figure 2.4 shows boxplots comparing needle lengths for the Aleppo pine and Torrey pine species of Examples 2.1 and 2.2.

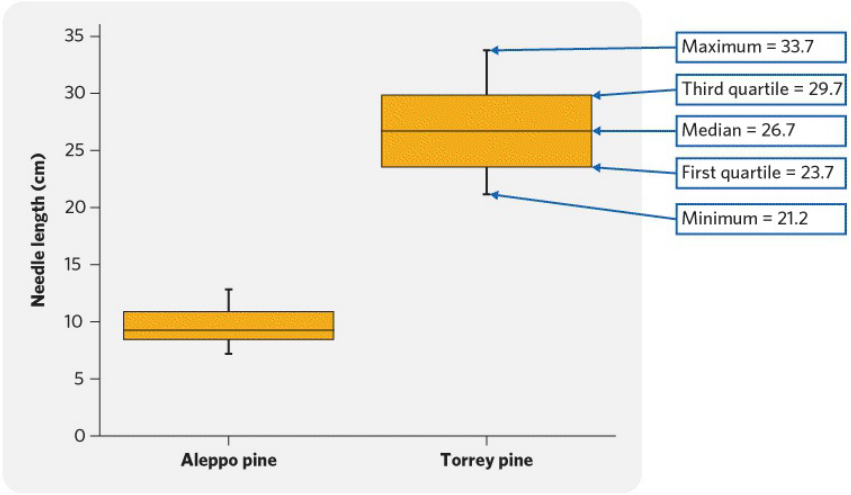


Figure 2.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.4 Boxplots comparing the lengths of 15 Aleppo pine needles and 18 Torrey pine needles.

BOXPLOT

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles Q1Q_1 and Q3.Q_3.
- A line in the box marks the median M.M.
- Lines extend from the box out to the smallest and largest observations.

Because boxplots show less detail than histograms and especially dotplots, they are best used for side-by-side comparison of more than one distribution, as in Figure 2.4. Be sure to include a numerical scale in the graph. When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The box shows the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set.

In a symmetric distribution, the first and third quartiles are equally distant from the median. In contrast, in most distributions that are skewed to the right, the third quartile will be farther above the median than the first quartile is below it. In most distributions that are skewed to the left, the first quartile will be farther below the median than the third quartile is above it. The extremes behave the same way, but remember that they are just single observations and may say little about the distribution as a whole.

EXAMPLE 2.8 Needle lengths of two pine tree species

In Examples 2.1 and 2.2 we examined sample sets of needles from two distinct species of pine trees, Aleppo and Torrey. In Figure 2.4 we can see that the needles in the Torrey pine set are all longer than the needles in the Aleppo pine set: The minimum length of the Torrey pine set is larger than the maximum length for the Aleppo pine set. Torrey pine needle lengths also have greater variability, as shown by the spread of the box and the spread between the extremes. Finally, the data for the Torrey pine are symmetrical, whereas the data for the Aleppo pine are mildly right-skewed.

CAUTION

Scientific publications sometimes provide summary statistics for two or more groups in the form of a graph displaying each mean with **error bars** extending on either side to show the standard deviation in each group. Another increasingly common display combines the raw data in the form of a dotplot with the group means shown as a symbol or the height of a bar and error bars scaled to equal one standard deviation on either side of the mean. Figure 2.5 shows such a graph for the needle lengths of the two species of pine trees. *Always check the legend of such graphs, because error bars may represent some other measure of variability* (such as the standard error of the mean or the margin of error, two concepts we will discuss in later chapters).

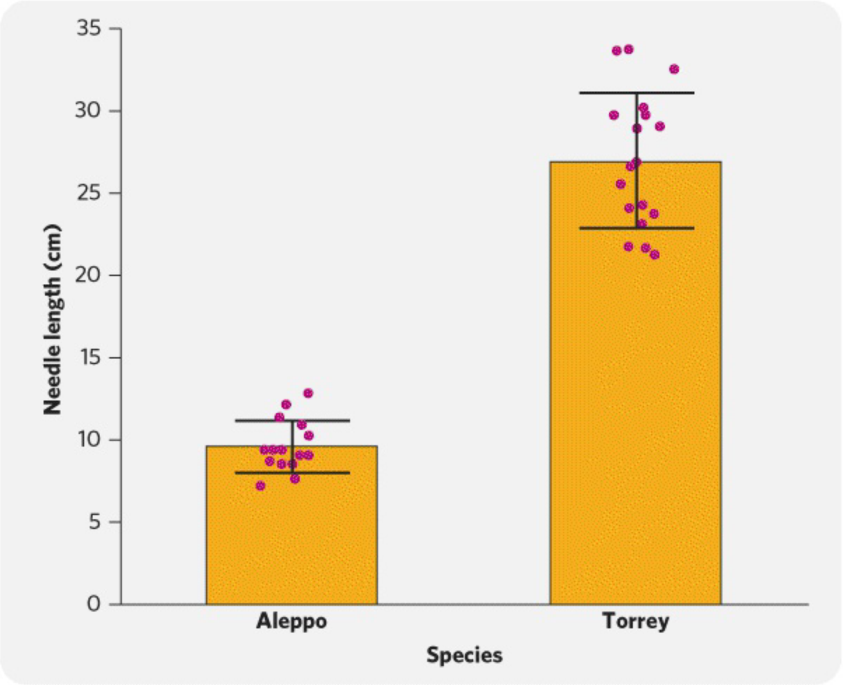


Figure 2.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.5 The same data as in Figure 2.4 now displayed as the mean (columns) plus and minus one standard deviation (error bars) overlaid on top of a dotplot of the raw data for both species.

APPLY YOUR KNOWLEDGE

2.6 Glucose levels. People with diabetes must monitor and control their blood glucose level. The goal is to maintain a “fasting plasma glucose” between approximately 90 and 130 milligrams per deciliter (mg/dl). Exercise 1.10 (page 21) gave the fasting plasma glucose levels for two groups of diabetics five months after they received either group instruction or individual instruction on glucose control.

a. Calculate the five-number summary for each of the two data sets.
b. Make side-by-side boxplots comparing the two groups (as in Figure 2.4). What can you say from this graph about the differences between the two diabetes control instruction methods?

c. Obtain the mean and standard deviation for each sample. Does this information give any clue about the shape of the two distributions?

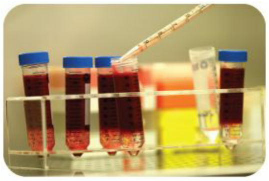
d. Add to the dotplots you created in Exercise 1.10 a symbol representing the mean of each group and error bars representing one standard deviation above and below the mean. You can do this by hand or with statistical software, if your software has this capability. Compare this graphical summary with the boxplot display you also created.

2.7 The cost of common blood tests. A 2014 study examined the prices (in dollars) billed by hospitals all over California for common blood tests such as a blood lipid panel (to check cholesterol level, first row) and a blood metabolic panel (including fasting plasma glucose level, second row). Here is how the findings were reported:[9]

N	Average	Standard deviation	Min	5th percentile	25th percentile	Median	75th percentile	per
178	299	759	10	76	134	220	303	
189	371	814	35	62	111	214	389	

a. Make side-by-side boxplots comparing the distributions of prices for the two procedures (as in Figure 2.4). Describe the distribution of prices for each blood test. Are there substantial differences between the two?

b. The report also provided the mean and standard deviation for the two distributions. Explain why in this case the mean and standard deviation would be poor choices of summary statistics to cite in a news report.



Claudio Gallone/AgeFotostock

2.8 $\bar{x}$ and s are not enough. The mean $\bar{x}$ and standard deviation s measure center and spread but are not a complete description of a distribution. Data sets with different shapes can have the same mean and standard deviation. To demonstrate this fact, use your calculator to find x¯ $\bar{x}$ and s s for these two small data sets. Then make a dotplot of each and comment on the shape of each distribution.

Data A	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data B	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

SPOTTING SUSPECTED OUTLIERS[i]

Figure 2.6 shows the dotplot of the sizes in cubic centimeters (cm^3) of 11 acorns from oak trees found on the Pacific coast.[10] The five-number summary for this distribution is

$$0.4 \qquad 1.6 \qquad 4.1 \qquad 6.0 \qquad 17.1$$

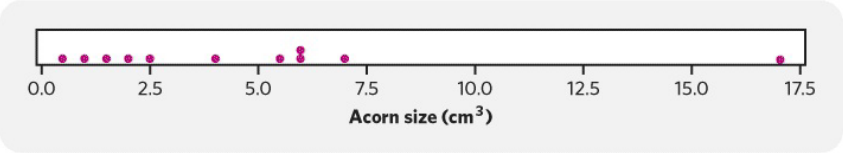


Figure 2.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.6 Dotplot of the sizes of 11 acorns.

How shall we describe the spread of this distribution? The dotplot shows that the largest observation is clearly an extreme that doesn't represent most of the data very well. The interquartile range (the range of the center half of the data) is a more resistant measure of spread. In this example, $\text{IQR}=6.0-1.6=4.4\text{ cm3}$.$IQR = 6.0 - 1.6 = 4.4\,\text{cm}^3$. However, the two sides of a skewed distribution have different spreads around the center of the distribution, so one number can't represent them both. The interquartile range is mainly used as the basis for a rule of thumb for identifying *suspected* outliers.

THE $1.5\times\text{IQR}$$1.5 \times IQR$ RULE FOR SUSPECTED OUTLIERS

Call an observation a *suspected* outlier if it falls more than $1.5\times\text{IQR}$$1.5 \times IQR$ above the third quartile or below the first quartile.

EXAMPLE 2.9 Spotting suspected outliers: acorns

For the acorn sizes data, the five-number summary is 0.4, 1.6, 4.1, 6.0, 17.1, so

$$1.5\times\text{IQR}=1.5\times 4.4=6.6.$$

$1.5 \times IQR = 1.5 \times 4.4 = 6.6.$

Any values not falling between

Q1-(1.5×IQR)=1.6-6.6=-5.0 andQ3+(1.5×IQR)=6.0+6.6=12.6

$$\begin{aligned} Q_1 - (1.5 \times IQR) &= 1.6 - 6.6 = -5.0 \text{ and} \\ Q_3 + (1.5 \times IQR) &= 6.0 + 6.6 = 12.6 \end{aligned}$$

are flagged as *suspected* outliers. In this case, the 1.5×IQR$1.5 \times IQR$ rule flags only one value in the data set, the largest value of 17.1, and suggests that it may be an outlier. Looking at the dotplot of acorn sizes in Figure 2.6, we can confirm that the largest acorn (17.1 cm^3) is indeed an outlier.

The 1.5×IQR$1.5 \times IQR$ rule is not a replacement for looking at the data. It is most useful when large volumes of data are scanned automatically. Some software programs create **modified boxplots,** which display any value outside of the 1.5×IQR$1.5 \times IQR$ interval around either quartile as individual data points (rather than including them within the long tail going to the minimum or maximum). Examine the data carefully and decide for yourself if a particular data point fits the definition of an outlier, “an individual value that falls outside the overall pattern.”

EXAMPLE 2.10 Spotting suspected outliers: flood events

In Example 2.4 we looked at the number of flood events recorded between 2009 and 2013 in each of 52 Atlantic coastal communities. The five-number summary for these data is 0, 5, 37.5, 73.5, 250 and

1.5×IQR=1.5×68.5=102.75

$$1.5 \times IQR = 1.5 \times 68.5 = 102.75$$

so that any values not falling between

Q1-(1.5×IQR)=5-102.75=-97.75 andQ3+(1.5×IQR)=73.5+102.75=176.25

$$\begin{aligned} Q_1 - (1.5 \times IQR) &= 5 - 102.75 = -97.75 \text{ and} \\ Q_3 + (1.5 \times IQR) &= 73.5 + 102.75 = 176.25 \end{aligned}$$

are flagged as *suspected* outliers. We can see from the dotplot in Figure 2.2 that the four largest values are flagged by this rule. Now we must decide whether they actually are outliers.

Figure 2.7 shows a combination histogram–modified boxplot created by the statistical software JMP. The four data points flagged by the 1.5×IQR$1.5 \times IQR$ rule are displayed as individual dots rather than as part of the boxplot's high whisker. A gap appears between these four points and the rest of the high whisker, but it is not very large. In fact, the histogram underneath has a very pronounced right skew without any gap. So are the *suspected* outliers flagged by the 1.5×IQR$1.5 \times IQR$ rule *actual* outliers? They are certainly somewhat more extreme than the rest of the distribution, yet they fit well within the overall right-skew pattern.

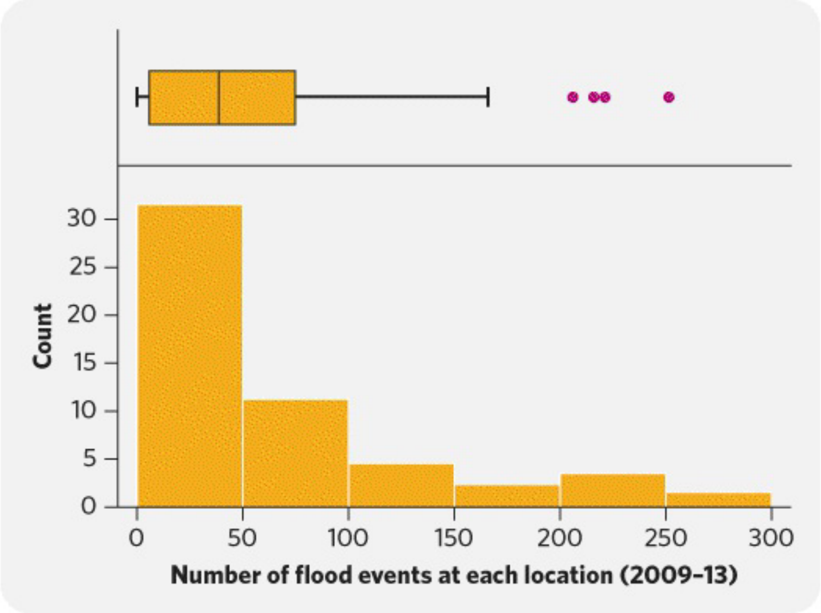


Figure 2.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 2.7 Histogram and modified boxplot (using the same horizontal scale) of the number of flood events recorded over the 2009-2013 period for each of 52 coastal communities.

Sometimes outliers are really obvious, as in Example 2.9, and sometimes they are more ambiguous, as in Example 2.10. It can be helpful to use words such as "mild" or "moderate" versus "extreme" to describe them. The four largest numbers of flood events may be described, for instance, as possibly mild outliers within a pronounced right skew. Simply be honest and describe your findings as they are.

APPLY YOUR KNOWLEDGE

2.9 Spider silk, continued. In Exercise 2.1 you plotted the silk yield stress for 21 female golden orb weaver spiders. Use the 1.5×IQR $1.5 \times IQR$ rule to identify suspected outliers.

2.10 Glucose levels, continued. In Exercise 2.6 you made side-by-side boxplots comparing the group and individual training methods for diabetes control. Use the 1.5×IQR$1.5 \times IQR$ rule to identify any suspected outliers. Then look at the raw data to determine if unusually high or low values in either data set actually are outliers.

CDC/Amanda Mills

DISCUSSION Dealing with outliers: recognition and treatment

When collecting data, we sometimes come across wild observations that clearly fall outside the overall pattern of data distribution. In this chapter, we have presented a way to identify suspected outliers and discussed how outliers can affect numerical summaries. But what are these wild observations and how do we deal with them? Entire books have been written on the topic. Here we describe three major types of wild observations and offer some general guidance on how to handle each type.

Human error in recording information

Errors in data recording are not that uncommon. Typos are an obvious concern. In addition, the data themselves are not always clean. Surveys of individuals are particularly prone to errors. People may forget, lie, or simply misunderstand a question. In an online survey, undergraduate students enrolled in a biostatistics course were asked to record their heights in inches. Of the 149 numerical values submitted, two wild observations appeared as 5.3 and 6. These are obvious errors. Maybe the students meant 5 feet 3 inches and 6 feet, respectively, or maybe the 5.3 value was a typo for 53 inches.

What should you do with wild observations that you have clearly identified as being errors in data entry? The obvious answer is that these values do not belong and should not be included with the whole data set. You might be able to correct the mistakes by checking your original records (notes, data tables, photos). Good scientific practice always includes keeping clear and extensive records of data and the way they were obtained.

Human error in experimentation or data collection

Sloppy experimentation methods can lead to unexpected results. If you forget to add bacteria to one of your Petri dishes and find that nothing grew in it, that’s just a silly mistake. But some experimental blunders lead to more interesting results.

Fleming, for instance, had less-than-ideal lab techniques. A few of his Petri dishes ended up being contaminated with a mold. Instead of simply throwing them away, Fleming noticed a halo around the mold where no bacteria grew. He went on to cultivate the mold and discover its antibiotic properties, revolutionizing medicine and later earning a Nobel Prize. When Pasteur was studying chicken cholera, his assistant left some bacterial cultures out while he went on vacation. The dried-out cultures failed to kill inoculated chickens, as other cultures usually did. The assistant’s first impulse was to discard the data, but Pasteur decided to take a closer look and explore the reason for this unexpected result. This work led him to understand the workings of the immune system and develop a vaccine.

Not all technical errors lead to fame or a Nobel Prize. Some mistakes are truly not worth a second look. Sometimes, however, the wild observations can be very interesting. You should always keep detailed notes of everything you do when gathering data, as this might help you identify mistakes and understand how they arose. Either way, these kinds of wild observations do not belong with the rest of your data (including them would be like comparing apples and oranges). If you suspect an experimental error occurred, the wild observation should be either ignored or studied separately.

Unexplainable but apparently legitimate wild observations

Many studies in the life sciences are conducted by collecting data about a small sample taken from the whole population of interest. In such cases, it can be difficult to determine whether a suspected outlier in a sample is truly a wild observation or just the consequence of studying only a small subset of the population. When you find a suspected outlier in a sample and have ruled out human error, you are faced with the challenging task of deciding what to do with it.

This is an important step because, for many statistical procedures, outliers are influential and can distort conclusions. Running the analysis first with the outlier and then again without the outlier, can help you determine whether the outlier affects your conclusions substantially or not. We will see in future chapters that some statistical methods are robust against mild outliers. That is, the method will be valid despite the presence of a mild outlier. Extreme outliers, however, are always a cause of concern. Many complex statistical approaches have been devised to deal with outliers, though most are well beyond the scope of this introductory textbook. Nonparametric tests, described in optional Chapter 27, are just a few examples. For simple studies, deciding what to do with a suspected outlier typically boils down to deciding whether to include the wild observation with the rest of the data.

To some extent, deciding whether to include or exclude wild observations in your analysis depends on the purpose of your study. Are you more interested in describing a population in its entire extent or only in its most typical, general pattern? For example, before Pasteur invented the pasteurization method to prevent the spoilage of liquids, wine and beer producers often had some batches that developed a troublesome sour taste. If you were studying the performance of various wine production methods, you would want to include bad batches in your analysis because they too reflect the performance of the method. In contrast, if you were studying the chemical composition of wines and beers to better understand their health effects, you would not include sour batches, because these would never be sold for consumption. Obviously, you should never discard wild observations simply because they lead to conclusions you do not like.

Lastly, and perhaps most importantly, whether you choose to keep or to discardthem, all outliers should be disclosed. Outliers can be very influential or have a tremendous scientific relevance, and you should not act as if they did not exist. You should also explain why you chose to keep or discard them. Honesty and full disclosure are the foundation of a good scientific methodology.

STATS IN YOUR WORLD

HOW EXTREME IS TOO EXTREME?

In 1949, Mr. Hadlum sought a divorce on the basis that his wife had given birth 349 days after his departure to war. He postulated that it was evidence of adultery because the average gestation time is 280 days. The judges ruled that such a long gestation time was unlikely but not impossible and denied his appeal. By comparison, while the average men’s height in the United States is 1.76 m (5’9”), the *Guinness Book of World Records* lists the tallest man in recorded history as a 2.72-m-tall American (8’11”).

STATS IN YOUR WORLD

HOW MANY WAS THAT?

An advocacy group claims, without much evidence, that 150,000 Americans have the eating disorder known as anorexia nervosa. Soon someone misunderstands and says that 150,000 people *die* from anorexia nervosa each year. This wild number gets repeated in countless books and articles. It really is a wild number: Only approximately 55,000 women age 15 to 44 (the main group affected by anorexia) die of *all causes* each year.

APPLY YOUR KNOWLEDGE

2.11 Blood pressure readings. A manufacturer of blood pressure monitoring devices lists 10 factors that can affect blood pressure readings:[11]

1. Blood pressure cuff too small
2. Blood pressure cuff used over clothing
3. Not resting for a few minutes
4. Arms, back, or feet unsupported
5. Emotional state
6. Talking
7. Smoking recently

8. Consuming alcohol or caffeine recently
9. Room temperature
10. Bladder full or empty

   a. Which factor or factors would lead to data points you would definitely want to discard in a study of blood pressure in healthy adults? Explain your reasoning. Which part of the Discussion on page 55 addresses this issue?
   b. Which factor or factors may lead to unusual blood pressure values that should not be ignored in a study of blood pressure in healthy adults? Explain your reasoning.
   c. If the manufacturer of a blood pressure cuff wanted to examine the variability of readings obtained at the doctor's office, should suspicious data points resulting from any of the 10 factors listed here be discarded? If your answer is different than the one you gave in part a, explain your reasoning. Which part of the Discussion on page 55 addresses this issue?

2.12 Temperatures records. In Exercise 1.14 (page 28) you downloaded historical temperature records for Los Angeles from NASA's website, and examined the distribution of meteorological annual mean temperatures ("metANN" column, in degrees Celsius, °C).

   a. Obtain the mean and the median for the metANN variable. How do they compare?
   b. Create a dotplot of these annual temperatures or refer to the one you created for Exercise 1.14. Several points with the value 999.9 stand out quite clearly on your dotplot. If you are not familiar with degrees Celsius, note that the human body's internal temperature is approximately 37°C. Is 999.9 a meteorologically plausible value for a mean annual temperature in Los Angeles? Which part of the Discussion on page 55 do these values most likely illustrate?
   c. Is it likely that NASA repeatedly made the same data entry error in recording temperatures in Los Angeles? What do you think is the most likely explanation for using the value 999.9 in the data

file you downloaded?

d. In Exercise 1.14. you cleaned up the data set so the data could be plotted properly. Obtain the mean and the median for the cleaned-up metANN variable. How do they compare now? Relate your findings to what you learned in this chapter.

ORGANIZING A STATISTICAL PROBLEM

Most of our examples and exercises have been aimed at helping you learn basic tools (graphs and calculations) for describing and comparing distributions. You have also learned principles that guide the use of these tools, such as “always start with a graph” and “look for the overall pattern and striking deviations from the pattern.” The data you work with are not just numbers: They describe specific settings, such as water salinity in the Everglades or needle length for two species of pine trees. Because data come from a specific setting, the final step in examining data is drawing a conclusion for that setting. For example, water depth in the Everglades has a yearly cycle that reflects Florida’s wet and dry seasons. Needle lengths are longer in Torrey pine trees than in Aleppo pine trees.

As you learn more statistical tools and principles, you will face more complex statistical problems. Although no framework can possibly accommodate all the varied issues arising in applying statistics to real settings, the following four-step thought process gives useful guidance. In particular, the first and last steps emphasize that statistical problems are tied to specific real-world settings and, therefore, involve more than doing calculations and making graphs.

ORGANIZING A STATISTICAL PROBLEM: THE FOUR-STEP PROCESS

STATE: What is the practical question, in the context of the real-world setting?

PLAN: What specific statistical operations does this problem call for?

SOLVE: Analyze the data with graphs and computations suitable for this problem.

CONCLUDE: Give your practical conclusion in the setting of the real-world problem.

To help you master the basics, many of our exercises will continue to tell you what todo—make a histogram, find the five-number summary, and so on. Of course, real-world statistical problems don’t come with detailed instructions. You will encounter some exercises that are more realistic, especially in the later chapters of the book. Use the four-step process as a guide to solving and reporting these problems. They are marked with the four-step icon, as the following example illustrates.

EXAMPLE 2.11 Highly Superior Autobiographical Memory

STATE: Some individuals have the ability to recall accurately vast amounts of autobiographical information without mnemonic tricks or extra practice. This ability is called Highly Superior Autobiographical Memory (HSAM). A study recruited 11 adults with confirmed HSAM and 15 control individuals of similar age without HSAM. All study participants were given a battery of cognitive and behavioral tests with the goal of finding out what might explain this extraordinary ability.[12]

First, autobiographical memory was assessed by asking each participant to recall in detail five important personal events. These events were selected by the researchers, and the participants did not know ahead of time which events they would have to recall. Answer accuracy was then verified from documents and interviews, and each correct detail was scored as one point. Individual total scores on this verifiable autobiographical memory task are displayed below.

HSAM	22	23	26	26	33	34	38	39	39	46	47				
Control	5	5	5	6	7	7	10	10	11	12	13	16	18	22	23

Do individuals with HSAM and without HSAM display distinct distributions of verifiable autobiographical memory scores? How do the mean scores compare?

PLAN: Use graphs and numerical descriptions to describe and compare these two distributions of verifiable autobiographical memory scores.

SOLVE: Dotplots work best for data sets of these sizes. Figure 2.8 displays stacked dotplots to facilitate comparison. The control group has a somewhat right-skewed distribution, so we might choose to compare the five-number summaries. But because the researchers plan to use $\bar{x}$ and s for further analysis, we instead calculate these measures:

Group	Mean score	Standard deviation
HSAM	33.9	8.8
Control	11.3	6.0

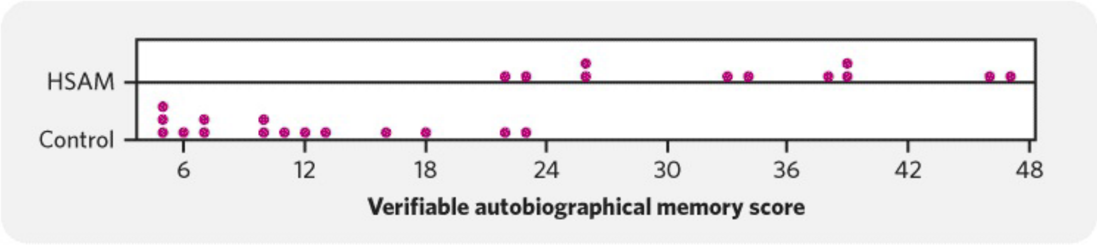


Figure 2.8

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.8 Dotplots comparing the distributions of scores on a test of verifiable autobiographical memory.

CONCLUDE: The two groups differ so much in verifiable autobiographical memory scores that there is little overlap among them. Overall, individuals with HSAM have higher scores than control individuals without HSAM. The mean scores are 33.9 for the HSAM group compared with only 11.3 for the control group, despite similar variation in individual scores (standard deviations 8.8 and 6.0, respectively). This first test confirms that individuals with HSAM have better recall of autobiographical events than ordinary individuals do. The researchers can now proceed to compare cognitive and behavioral traits in both groups.

APPLY YOUR KNOWLEDGE

2.13 Highly Superior Autobiographical Memory, continued. The study in Example 2.11 assessed common obsessional symptoms in the HSAM and the control individuals, as previous findings had suggested a possible obsessional component to HSAM.

The study participants completed the Leyton Obsessional Inventory test, which has a maximum score of 30 points. The results for the 11 HSAM individuals and for 14 of the controls are displayed below (one individual in the control group left before completing this test).

HSAM	2	4	5	8	8	9	9	10	11	11	12			
Control	1	2	2	2	2	4	4	4	5	7	8	8	12	12

How do individuals with and without HSAM compare in terms of obsessional score? Follow the four-step process in reporting your work.

CHAPTER 2 SUMMARY

- A numerical summary of a quantitative distribution should report at least its **center** and its **spread** or **variability.**
- The **mean** x̄$\bar{x}$ and the **median M**M describe the center of a distribution in different ways. The mean is the arithmetic average of *all* the observations, and the median is the midpoint of the values.
- When you use the median to indicate the center of the distribution, describe the spread by giving the **five-number summary** consisting of the median, the quartiles, and the smallest and largest individual observations. The **first quartile** Q_1 has one-quarter of the observations below it, and the **third quartile** Q_3 has three-quarters of the observations below it.
- When you use the mean to indicate the center of a sample distribution, describe the spread by giving the **standard deviation s.**s. The standard deviation is the square root of the **variance** s2.s^2. They are both zero when there is no spread and become larger as the spread increases.
- A **resistant measure** of any aspect of a quantitative distribution is relatively unaffected by changes in the numerical value of a small proportion of the total number of observations, no matter how large these changes are.
- The median and quartiles are resistant, but the mean and standard deviation are not. The mean and standard deviation are good descriptions for symmetric distributions without outliers. The five-number summary is a better description for skewed distributions. However, numerical summaries do not fully describe the shape of a distribution. Always plot your data.
- **Boxplots** based on the five-number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the distribution. The median is marked within the box. Lines extend from the box to the extremes and show the full spread of the data.

- A statistical problem has a real-world setting. You can organize many problems by using the four-step method: **State, Plan, Solve,** and **Conclude.**

THIS CHAPTER IN CONTEXT

In this chapter we have continued our study of exploratory data analysis. Numerical summaries can be useful for describing a single distribution as well as for comparing the distributions of several groups of observations.

Two important features of a quantitative distribution are its center and its spread. The mean and standard deviation are excellent numerical summaries for distributions that are approximately symmetric without outliers. If a distribution is not symmetric, has outliers, or both, the five-number summary provides a better, more comprehensive description. But if a distribution is complex, with clusters or multiple peaks, for instance, reducing the information of the distribution to a few numbers would be misleading. This is one reason why data should always be graphed before computing and communicating numerical summaries.

We will see later, starting with Chapter 14, how numerical summaries from random samples are used for estimating population parameters or for comparing several population distributions. For this reason, it is important to distinguish the notation for the mean $\bar{x}$ $\bar{x}$ and standard deviation **s** s of a sample from the notation for the mean and standard deviation of a population (μ μ and σ, σ, respectively).

CHECK YOUR SKILLS

2.14 Cesium-137 is a waste product of nuclear reactors. A study examined the cesium-137 tissue concentration of a random sample of 15 Pacific bluefin tuna, *Thunnus orientalis*, captured off the coast of California four months after the Fukushima (Japan) nuclear reactor meltdown of 2011. Here are the findings, in becquerels per kilogram of dry tissue:[13]

4.8 7 6.2 7.3 6.0 7.3 3.7 9.4 8.4 6.0 5.0 5.1 6.0 7.3 4.7

What is the mean of these data?

a. 5.70
b. 6.00
c. 6.28

2.15 What is the median of the data in Exercise 2.14?

a. 5.70
b. 6.00
c. 6.28

2.16 (Optional) What is the interquartile range, IQR,IQR, of the data in Exercise 2.14?

a. 1.3
b. 2.0
c. 2.3

2.17 Ifa distribution is clearly skewed to the right,

a. the mean is less than the median.
b. the mean and the median are equal.
c. the mean is greater than the median.

2.18 What percent of the observations in a distribution lie between the first quartile and the third quartile?

a. 25%
b. 50%

c. 75%

2.19 To make a boxplot of a distribution, you must know

a. all the individual observations.
b. the mean and the standard deviation.
c. the five-number summary.

2.20 What are all the values that a standard deviation ss can possibly take?

a. 0≤s$0 \leq s$
b. 0≤s≤1$0 \leq s \leq 1$
c. -1≤s≤1$-1 \leq s \leq 1$

2.21 What is the approximate value of the standard deviation of the 15 cesium-137 values in Exercise 2.14? (Use your calculator.)

a. 1.47
b. 1.52
c. 2.32

2.22 What are the correct units for the standard deviation of the 15 cesium-137 values in Exercise 2.14?

a. no units—it's just a number
b. becquerels per kilogram of dry tissue
c. becquerels squared per square kilogram of dry tissue

2.23 Which of the following is least affected if an extreme high outlier is added to your data?

a. the median
b. the mean
c. the standard deviation

CHAPTER 2 EXERCISES

2.24 Food oils and health. Table 1.2 (page 34) gives the ratio of omega-3 to omega-6 fatty acids in common food oils. Exercise 1.36 asked you to plot the data. The distribution is strongly right-skewed with a high outlier. Do you expect the mean to be greater than the median, about equal to the median, or less than the median? Why? Calculate $\bar{x}$ and M and verify your expectation.

2.25 Lyme disease. Figure 1.8 (page 19) shows the age in years of 241,931 patients diagnosed with Lyme disease in the United States between 1992 and 2006. Give a brief description of the important features of the distribution. Explain why no numerical summary would appropriately describe this distribution.

2.26 Making resistance visible. In the *Mean and Median* applet, place three observations on the line by clicking below it: two close together near the center of the line, and one somewhat to the right of these two.

a. Pull the single rightmost observation out to the right. (Place the cursor on the point, hold down a mouse button, and drag the point.) How does the mean behave? How does the median behave? Explain briefly why each measure acts as it does.
b. Now drag the single rightmost point to the left as far as you can. What happens to the mean? What happens to the median as you drag this point past the other two (watch carefully)?

2.27 Anorexia nervosa. Figure 1.9 (page 20) is a histogram of the distribution of age at onset of anorexia for 691 Canadian girls diagnosed with the disorder. If you round the age to whole numbers of years, the first bar of the histogram (the first class) would include all girls diagnosed during their 11th year. With a little care, you canfind the median and the quartiles from the histogram. What are these numbers? How did you find them?

2.28 Laughing behavior. A study of freely forming groups in bars throughout Europe examined the number of individuals found in groups whose members were laughing together. The study reported on a total of 501 laughing groups, distributed as follows:[14]

Number of individuals in group	2	3	4	5	6
Number of groups	254	168	52	21	6

a. Obtain the five-number summary for these data and display the results in a hand-drawn boxplot. Describe the distribution of the number of individuals in the freely forming laughing groups examined by this study.

b. Based on your boxplot, do you expect the mean laughing group size to be smaller, similar, or larger than the median group size? Explain your reasoning.

c. Compute the mean laughing group size, paying attention to the fact that the data are presented in frequencies (see Exercise 2.3 for reference). Does the computed mean fit your expectation as stated in part b?

2.29 Metabolic rate. In Example 2.7 you examined the metabolic rates of 7 men. Here are the metabolic rates for 12 women from the same study:

995 1425 1396 1418 1502 1256 1189
913 1124 1052 1347 1204

a. The most common methods for formal comparison of two groups use $\bar{x}$ $\bar{x}$ and ss to summarize the data. What kinds of distributions are best summarized by $\bar{x}$$\bar{x}$ and s?s?

b. Make a summary graph comparing the metabolic rates of the 7 men and 12 women, as in Figure 2.4. What can you conclude about these two groups from your graph?

2.30 Behavior of the median. Place five observations on the line in the *Mean and Median* applet by clicking below it.

a. Add one additional observation *without changing the median*. Where is your new point?

b. Use the applet to convince yourself that when you add yet another observation (there are now seven in all), the median does not change no matter where you put the seventh point. Explain why this must be true.

2.31 Nanomedicine. In Exercise 1.41 (page 36) you graphed the distribution of ovarian tumor increases under two experimental conditions: a new nanoparticle-based delivery system for a suicide gene therapy or an inactive buffer solution.

a. Make a boxplot comparing tumor increase under the two conditions and compute the mean and standard deviation for each condition.

b. Write a short description of the experimental results based on your work in a.

2.32 Guinea pig survival times. Here are the survival times (in days) of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment:[15]

43	45	53	56	56	57	58	66	67	73	74	79
80	80	81	81	81	82	83	83	84	88	89	91
91	92	92	97	99	99	100	100	101	102	102	102
103	104	107	108	109	113	114	118	121	123	126	128
137	138	139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511	522	598

a. Plot and describe the distribution of survival times. As is often the case with survival times, this distribution is strongly skewed to the right.

b. Which numerical summary would you choose for these data? Explain why.

c. Obtain your chosen summary. How does it reflect the skewness of the distribution?

2.33 A standard deviation contest. This is a standard deviation contest. You must choose four numbers from the whole numbers 0 to 10, with repeats allowed.

a. Choose four numbers that have the smallest possible standard deviation.
b. Choose four numbers that have the largest possible standard deviation.
c. Is more than one choice possible in either a or b? Explain.

2.34 Does breastfeeding weaken bones? Breastfeeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral content. Researchers compared 47 breastfeeding women with 22 women of similar age who were neither pregnant nor lactating. They measured the percent change in the mineral content of the women's spines over three months. A negative value indicates a loss in mineral content. Here are the data:[16]

Breastfeeding women

−4.7	−2.5	−4.9	−2.7	−0.8	−5.3
−8.3	−2.1	−6.8	−4.3	2.2	−7.8
−3.1	−1.0	−6.5	−1.8	−5.2	−5.7
−7.0	−2.2	−6.5	−1.0	−3.0	−3.6
−5.2	−2.0	−2.1	−5.6	−4.4	−3.3
−4.0	−4.9	−4.7	−3.8	−5.9	−2.5
−0.3	−6.2	−6.8	1.7	0.3	−2.3
0.4	−5.3	0.2	−2.2	−5.1	

Other women

2.4	0.0	0.9	−0.2	1.0	1.7
2.9	−0.6	1.1	−0.1	−0.4	0.3
1.2	−1.6	−0.1	−1.5	0.7	−0.4
2.2	−0.4	−2.2	−0.1		

Do the data show distinctly greater bone mineral loss among the breastfeeding women? Be sure to consider in your interpretation the understanding that a bone mineral loss is reflected by a negative value here. Follow the four-step process illustrated by Example 2.11.

2.35 Cicadas as fertilizer? Every 17 years, swarms of cicadas emerge from the ground in the eastern United States, live for about six weeks, then die. (There are several "broods," so we see cicada eruptions more often than every 17 years.) So many cicadas die that their bodies may serve as fertilizer and increase plant growth. In an experiment, a researcher added 10 cicadas under some plants in a natural plot of American bellflowers in a forest, leaving other plants undisturbed. One of the variables studied was the size of seeds produced by the plants. Here are data (seed mass in milligrams) for 39 plants fertilized with cicadas and 33 undisturbed plants:[17]

Cicada plants				Undisturbed plants			
0.237	0.277	0.241	0.142	0.212	0.188	0.263	0.253
0.109	0.209	0.238	0.277	0.261	0.265	0.135	0.170
0.261	0.227	0.171	0.235	0.203	0.241	0.257	0.155
0.276	0.234	0.255	0.296	0.215	0.285	0.198	0.266
0.239	0.266	0.296	0.217	0.178	0.244	0.190	0.212
0.238	0.210	0.295	0.193	0.290	0.253	0.249	0.253
0.218	0.263	0.305	0.257	0.268	0.190	0.196	0.220
0.351	0.245	0.226	0.276	0.246	0.145	0.247	0.140
0.317	0.310	0.223	0.229	0.241			
0.192	0.201	0.211					

Do these data support the idea that dead cicadas can serve as fertilizer? Follow the four-step process in your work.

2.36 Daily activity and obesity. People gain weight when they take in more energy from food than they expend. Table 2.1 compares volunteer subjects who were lean with others who were mildly obese. None of the subjects followed an exercise program. The subjects wore sensors that recorded every move for 10 days. The table shows the average minutes per day spent in activity (standing and walking) and in lying down.[18] Compare the distributions of time spent actively for lean and obese subjects as well as the distributions of time spent lying down. How does the behavior of lean and mildly obese people differ? Follow the four-step process in your work.

TABLE 2.1 Time (minutes per day) active and lying down by lean and obese subjects

	Lean subjects			Obese subjects	
Subject	**Stand/Walk**	**Lie**	**Subject**	**Stand/Walk**	**Lie**
1	511.100	555.500	11	260.244	521.044
2	607.925	450.650	12	464.756	514.931
3	319.212	537.362	13	367.138	563.300
4	584.644	489.269	14	413.667	532.208
5	578.869	514.081	15	347.375	504.931
6	543.388	506.500	16	416.531	448.856
7	677.188	467.700	17	358.650	460.550
8	555.656	567.006	18	267.344	509.981
9	374.831	531.431	19	410.631	448.706
10	504.700	396.962	20	426.356	412.919

2.37 Logging in the rainforest. "Conservationists have despaired over destruction of tropical rainforest by logging, clearing, and burning." These words begin a report on a statistical study of the effects of logging in Borneo.[19] Researchers compared forest plots that had never been logged (Group 1) withsimilar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). All plots were 0.1 hectare in area. Here are the counts of trees for plots in each group:

Group 1	27	22	29	21	19	33	16	20	24	27	28	19
Group 2	12	12	15	9	20	18	17	14	14	2	17	19
Group 3	18	4	22	15	18	19	22	12	12			

To what extent has logging affected the count of trees? Follow the four-step process in reporting your work.

Exercises 2.38 to 2.41 make use of the optional material on the 1.5×IQR $1.5 \times IQR$ *rule for suspected outliers.*

2.38 Mercury levels, continued. In Exercise 2.4 you obtained the five-number summary of the distribution of blood mercury levels among 4134 pregnant British women enrolled in a scientific study.

a. Use the 1.5×IQR$1.5 \times IQR$ rule to determine whether the smallest and the largest values quoted qualify as suspected outliers.

b. The published quote included a number of percentiles besides those making up the five-number summary. Do any of the quoted values beside the minimum and the maximum qualify as suspected outliers based on the 1.5×IQR$1.5 \times IQR$ rule?

c. Based on your analysis in parts a and b, does the distribution of blood mercury levels in this study include at least one actual outlier or is it simply skewed? Explain your reasoning.

2.39 Aggression and social status in macaque monkeys. The boxplots in Figure 2.9 summarize the number of aggressive behaviors directd at female macaque monkeys of differing social status living in experimentally arranged groups.[20] These are modified boxplots that indicate suspected outliers by a dot, using the 1.5×IQR$1.5 \times IQR$ rule. Based on these plots, describe how “aggressions received” by female macaque monkeys vary depending on their dominance rank within the group.

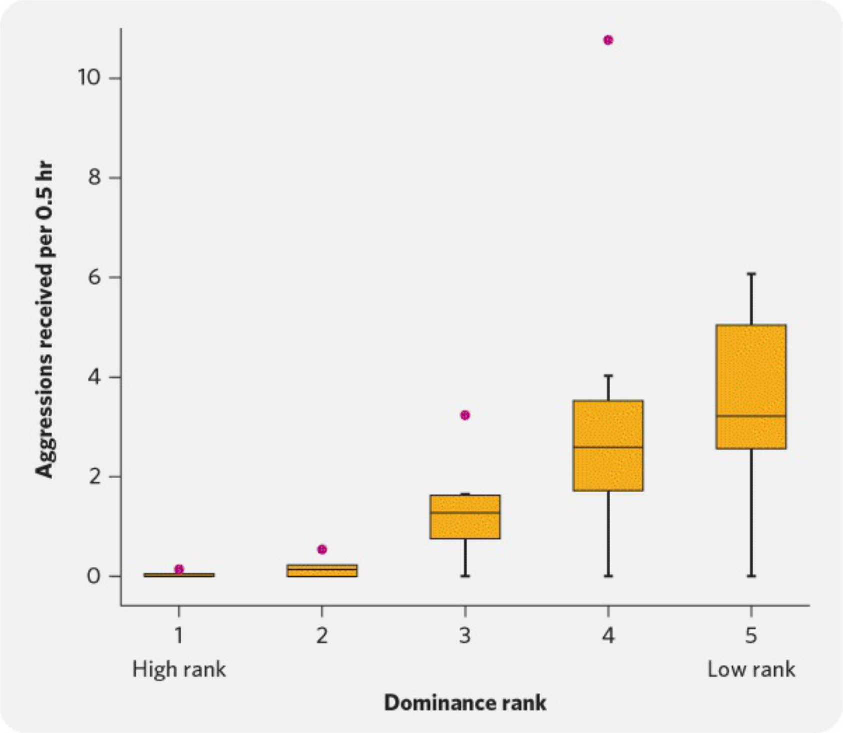


Figure 2.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 2.9 Boxplots of the distributions of number of aggressions received per half hour, for female macaque monkeys of differing social status.

2.40 Golden orb weavers. Exercise 2.1 described a study of female golden orb weaver spiders. The study also reported the body mass (in grams) for each of the 21 spiders. Here are the data:

0.04	0.11	0.16	0.07	0.13	0.1	0.17
0.25	0.36	0.33	0.29	0.14	0.32	0.57
0.31	0.79	0.49	0.64	0.6	0.99	0.81

a. Give the five-number summary, mean, and standard deviation of this distribution. How do the mean and the median compare?

b. Does the 1.5×IQR$1.5 \times IQR$ rule identify any suspected outliers?

c. What do the results in parts a and b suggest about the distribution of body mass in female golden orb weavers? Make a dotplot to confirm your interpretation.

2.41 Young Americans. The dotplot in Figure 1.16 (page 31) displays the distribution of the percents of individuals younger than age 18 in each of the 50 states and the District of Columbia. Note that the values in the dotplot are reported in increments of 0.5 percentage points.

a. Give the five-number summary of this distribution.

b. Does the 1.5×IQR$1.5 \times IQR$ rule identify the minimum (District of Columbia) and maximum (Utah) as suspected outliers? Does it also flag any other states?

c. The mean for these 51 values is 23.755. How does it compare with the median? Explain how this is possible.

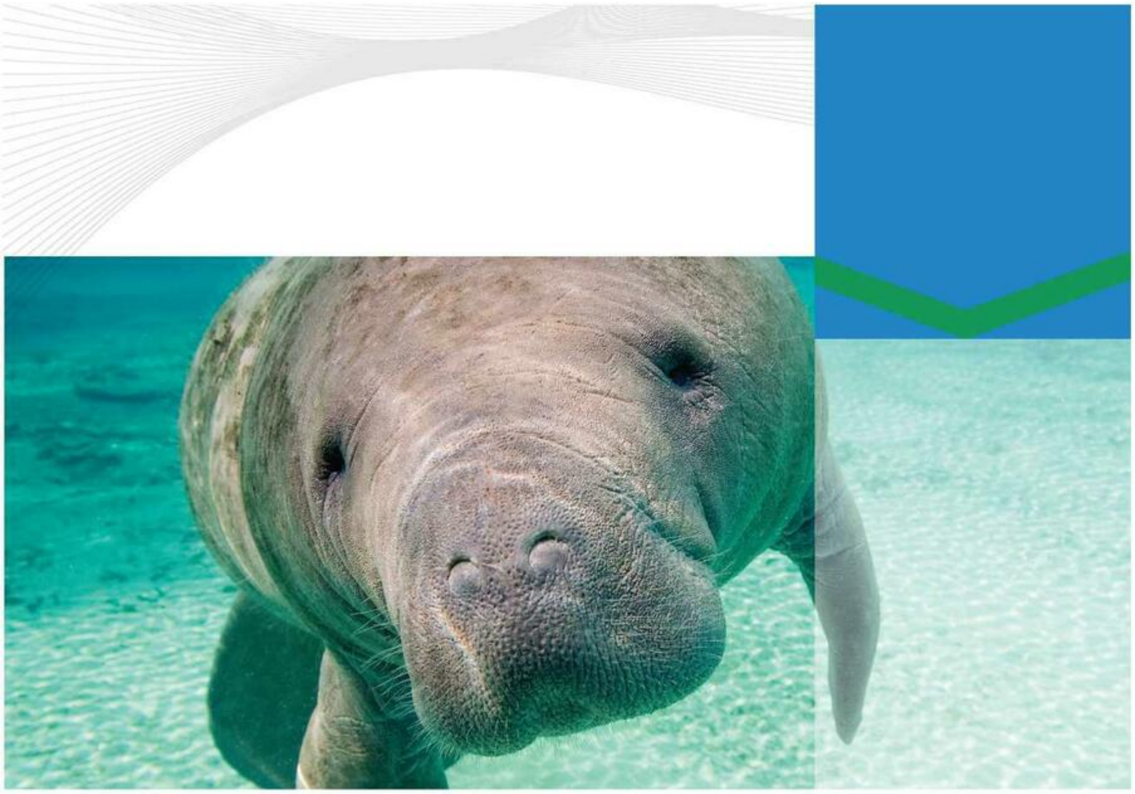
Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Note that short answers are not available for these exercises so that they can be used for class work or assignments.*

2.42 Elderly health. A study examined the medical records of elderly patients to determine whether there are differences between men and women in their calcium or inorganic phosphorus blood levels (both in millimoles per liter, mmol/l). The data file *Large.Calcium* contains the data.[21]

LARGE DATA SET

a. Make side-by-side boxplots comparing the calcium levels for men (sex=1)$(sex = 1)$ and for women (sex=2).$(sex = 2)$. Do the same for the inorganic phosphorus levels. What do you conclude?
b. Another purpose of the study was to determine whether different laboratories yield substantially different results. Use boxplots to compare the calcium levels obtained by the six different labs. Do you see notable differences? Do the same for the inorganic phosphorus levels. Note that the labs had each received blood samples from a different set of patients. How does that affect your interpretation of the boxplots?

CHAPTER 3 Scatterplots and Correlation



Wayne Lynch/All Canada/Shutterstock.com

IN THIS CHAPTER WE COVER...

- Explanatory and response variables
- Relationship between two quantitative variables: scatterplots
- Adding categorical variables to scatterplots
- Measuring linear association: correlation

A medical study finds that lung capacity decreases with the number of cigarettes smoked per day. The Department of Motor Vehicles warns that alcohol consumption reduces reflex time and that the effect is larger as more alcohol is consumed. These and many other statistical studies look at the *relationship between two variables.* Statistical relationships are overall tendencies, not iron-clad rules. They allow individual exceptions.

Although smokers, on average, die younger than nonsmokers, some people live to age 90 while smoking three packs a day.

To understand a statistical relationship between two variables, we measure both variables on the same individuals. Often, we must examine other variables as well. To conclude that smoking reduces lung capacity, for example, the researchers had to eliminate the effect of other variables such as each person’s size and exercise habits. In this chapter we begin our study of relationships between variables. One of our main themes is that the relationship between two variables can be strongly influenced by other variables that are lurking in the background.

EXPLANATORY AND RESPONSE VARIABLES

We think that blood alcohol content helps explain variations in reflex time and that smoking influences lung capacity and, eventually, life expectancy. In each of these relationships, the two variables play different roles: One explains or influences the other.

RESPONSE VARIABLE, EXPLANATORY VARIABLE

A **response variable** measures an outcome of a study. An **explanatory variable** may explain or influence changes in a response variable.

You will often find explanatory variables called **independent variables,** and response variables called **dependent variables.** The idea behind this language is that the response variable depends on the explanatory variable. Because "independent" and "dependent" have other meanings in statistics that are unrelated to the explanatory–response distinction, we prefer to avoid those words in this textbook.

It is easiest to identify explanatory and response variables when we actually set values of one variable to see how it affects another variable.

EXAMPLE 3.1 Beer and blood alcohol

How does drinking beer affect the level of alcohol in a person's blood? The legal blood alcohol content limit for driving is now 0.08% in all states. Student volunteers at Ohio State University drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol content. Number of beers consumed is the explanatory variable, and percent alcohol in the blood is the response variable.

When we don't set the values of either variable but simply observe both variables, there may or may not be explanatory and response variables, depending on how we plan to use the data.

EXAMPLE 3.2 Height

The National Center for Health Statistics (NCHS) surveys the American population and collects, for each individual in the survey, information about body height, age, sex, and a long list of other attributes. The purpose of the NCHS survey is to document the characteristics of the American population. There is no explanatory or response variable in this context.

A pediatrician looks at the same data with an eye toward using age and sex, along with other variables such as ethnicity, to discuss a child's growth. Now age and sex are explanatory variables, and height is the response variable.

In many studies, the goal is to show that changes in one or more explanatory variables actually *cause* changes in a response variable. Many explanatory–response relationships, however, do not involve direct causation. The age and sex of a child can help predict future height, but they certainly do not *cause* a particular height.

Most statistical studies examine more than one variable. Fortunately, statistical analysis of several-variable data builds on the tools we used to examine individual variables. The principles that guide our work also remain the same:

- Plot your data. Look for overall patterns and deviations from those patterns.
- Based on what your plot shows, choose numerical summaries for some aspects of the data.

APPLY YOUR KNOWLEDGE

3.1 Explanatory and response variables? In each of the following situations, is it more reasonable to simply explore the relationship between the two variables or to view one of the variables as an explanatory variable and the other as a response variable? In the latter case, which is the explanatory variable and which is the response variable?

a. The typical amount of calories a person consumes per day and that person’s percent of body fat
b. The weight in kilograms and height in centimeters of a person
c. Inches of rain in the growing season and the yield of corn in bushels per acre
d. A person’s leg length and arm length, in centimeters

3.2 Coral reefs. How sensitive to changes in water temperature are coral reefs? To find out, scientists examined data on summer sea-surface temperatures and coral growth per year at locations in the Red Sea.[1] What are the explanatory and response variables? Are they categorical or quantitative?

3.3 Global climate change. In 2015, the United Nations signed a new global climate change agreement in Paris. Around that time, the Pew Research Center conducted an international survey to explore the relationship between the level of concern over climate change and CO_2 emissions per capita in each country.[2]

a. Does this study have a clear response variable? Would it be possible for a country’s per-capita CO_2 emissions to influence its citizens’ level of concern over global climate change? Would it be possible for people’s level of concern over global climate change to influence a country’s per-capita CO_2 emissions? Explain your reasoning.
b. The stated objective of the survey was to examine factors that may help explain opinions about global climate change worldwide. In this context, which of the two variables cited would be the explanatory variable?

RELATIONSHIP BETWEEN TWO QUANTITATIVE VARIABLES: SCATTERPLOTS

A **bivariate** analysis, examining the relationship between two variables, begins by plotting the data. The most useful graph for displaying the relationship between two quantitative variables is a *scatterplot.*

SCATTERPLOT

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

Always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot. As a reminder, we usually call the explanatory variable x and the response variable y. If there is no explanatory–response distinction, either variable can go on the horizontal axis.

EXAMPLE 3.3 An endangered species: the manatee

Manatees are large, herbivorous, aquatic mammals found primarily in the rivers and estuaries of Florida. This endangered species suffers from cohabitation with human populations, and many manatees die each year from collisions with powerboats. Following our four-step process (page 58), let's look at the influence of the number of powerboats registered on manatee deaths from collisions with powerboats. We examine the relationship between the number of manatee deaths from powerboat collisions and the number of powerboats registered in any given year between 1977 and 2016, based on the data reported in Table 3.1.[3]

TABLE3.1 Powerboat registrations (in thousands) and manatee deaths from Florida

Year	Powerboats	Deaths	Year	Powerboats	Deaths	Year	Powerboats	Deat
1977	447	13	1987	645	39	1997	755	54
1978	460	21	1988	675	43	1998	809	66
1979	481	24	1989	711	50	1999	830	82
1980	498	16	1990	719	47	2000	880	78
1981	513	24	1991	681	55	2001	944	81
1982	512	20	1992	679	38	2002	962	95
1983	526	15	1993	678	35	2003	978	73
1984	559	34	1994	696	49	2004	983	69
1985	585	33	1995	713	42	2005	1010	79
1986	614	33	1996	732	60	2006	1024	92

Wayne Lynch/All Canada /Shutterstock.com

STATE: The number of powerboats registered in Florida varies from year to year. Does this variable help explain the differences from year to year in the number of manatee deaths from collisions with powerboats?

PLAN: Examine the relationship between powerboats registered and manatee deaths from collision. Choose the explanatory and response variables (if any). Make a scatterplot to display the relationship between the variables. Interpret the plot to understand the relationship.

SOLVE (FIRST STEPS): We suspect that “powerboats registered” will help explain “manatee deaths from collisions.” So “powerboats registered” is the explanatory variable, and “manatee deaths from collisions” is the response variable. Time (the year in which the data were gathered) is not a variable of interest here. Instead, we want to see how the variable manatee deaths changes when the variable powerboat registrations changes, so we put powerboat registrations (the explanatory variable, expressed in thousands) on the horizontal axis. Figure 3.1 is the scatterplot of these data. Each point represents a single year. In 1997, for example, there were 755,000 powerboats registered and 54 manatee deaths due to powerboat collisions. Find 755 on the xx (horizontal) axis and 54 on the yy (vertical) axis. In Figure 3.1 the year 1997 appears as the point (755, 54) above 755 and to the right of 54.

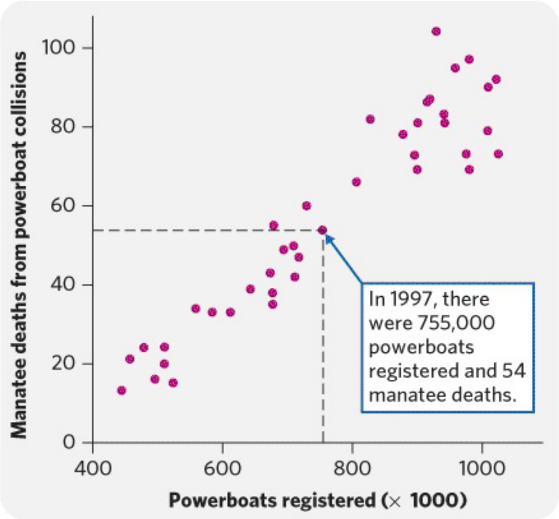


Figure 3.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 3.1 Scatterplot of the number of manatee deaths due to powerboat collisions in Florida each year against the number of powerboats registered (in thousands) that same year. The dotted lines intersect at the point (755, 54), the data for year 1997.

A scatterplot is meant to examine a bivariate relationship. To achieve that purpose, the graph should zoom in on the pattern of the relationship, avoiding extra blank space on the edges, and give equal emphasis to both axes, resizing the plot to be roughly square rather than rectangular in shape. Taking the time to properly scale a scatterplot is important because *our eyes can be fooled by changing the plotting scales or the amount of empty space around the cloud of points.*[4]

CAUTION

To interpret a scatterplot, apply the strategies of data analysis learned in Chapters 1 and 2.

EXAMINING A SCATTERPLOT

In any graph of data, look for the **overall pattern** and for striking **deviations** from that pattern.

You can describe the overall pattern of a scatterplot by the

direction, form, and **strength** of the relationship.

An important kind of deviation is an **outlier,** an individual value that falls outside the overall pattern of the relationship.

EXAMPLE 3.4 An endangered species: the manatee

SOLVE (INTERPRET THE PLOT): Figure 3.1 shows a clear *direction:* The overall pattern moves up, from lower left to upper right. That is, years in which powerboat registrations were higher tend to have higher counts of manatee deaths from collisions. We call this sort of pattern a positive association between the two variables. The *form* of the relationship is **linear;** that is, the overall pattern follows a straight line from lower left to upper right. The *strength* of a relationship in a scatterplot is determined by how closely the points follow a clear form. The overall relationship in Figure 3.1 is strong because the points are fairly close to forming a straight line.

CONCLUDE: The number of powerboat registrations explains much of the variation in the number of manatee deaths from collisions. Years that had fewer powerboats registered also tended to have fewer accidental manatee deaths. To preserve this endangered species, restricting the number of powerboats registered might be helpful. However, the scatterplot in Figure 3.1 does not take into consideration other factors, such as speed limits, fines, or driver education, that might also influence the number of manatee deaths from collisions. Our conclusions are limited to the available data and what they say about the relationship between powerboat registrations and manatee accidents.

POSITIVE ASSOCIATION, NEGATIVE ASSOCIATION

Two variables are **positively associated** when above-average values of one variable tend to accompany above-average values of the other, and below-average values also tend to occur together.

Two variables are **negatively associated** when above-average values of one variable tend to accompany below-average values of the other, and vice versa.

EXAMPLE 3.5 Does fidgeting keep you slim?

Obesity is a growing problem around the world. Surprisingly, some people don't gain weight even when they overeat. Perhaps fidgeting and other "nonexercise activity" (NEA) explains why—some people may spontaneously increase their nonexercise activity when fed more. Researchers deliberately overfed 16 healthy young adult volunteers for 8 weeks. They measured fat gain (in kilograms, kg) and, as an explanatory variable, change in energy use (in kilocalories, or Calories, Cal) from activity other than deliberate exercise—fidgeting, daily living, and the like. Here are the data:[5]

NEA change (Cal)	−94	−57	−29	135	143	151	245	355
Fat gain (kg)	4.2	3.0	3.7	2.7	3.2	3.6	2.4	1.3
NEA change (Cal)	392	473	486	535	571	580	620	690
Fat gain (kg)	3.8	1.7	1.6	2.2	1.0	0.4	2.3	1.1

Robert Daly/Getty Images

Do people with larger increases in NEA tend to gain less fat? To see the pattern in the data, we make a scatterplot with NEA on the horizontal axis (the explanatory variable), as displayed in Figure 3.2.

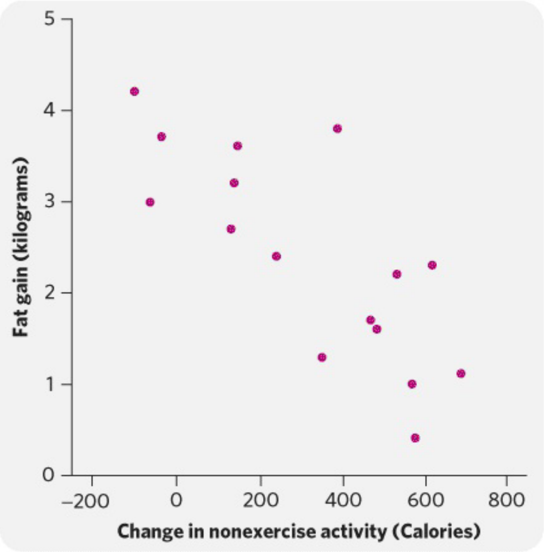


Figure 3.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 3.2 Scatterplot of fat gain after 8 weeks of overeating plotted against increase in nonexercise activity over the same period.

The plot shows a negative association: People with larger increases in NEA tend to gain less fat. The form of the association between NEA and fat gain is linear. The association is moderately strong because the points make up a linear pattern but deviate quite a lot from the line.

When examining the relationship between two quantitative variables, we need to consider outliers in the context of this relationship. That is, we need to look for points on the scatterplot that do not fit the overall pattern, regardless of whether they would be identified as outliers in a dotplot or histogram of each distribution separately. What to do about outliers is as challenging in the context of a bivariate analysis as it was in the univariate situation of Chapter 2.

EXAMPLE 3.6 Magellanic penguins

Punta Tombo, Argentina, is home to the world’s largest breeding colony of Magellanic penguins (*Spheniscus magellanicus*). Researchers have studied this colony since 1983, recording each year how many eggs were laid, how many chicks hatched, how many died of various causes, and how many successfully left the nest (fledged). Figure 3.3 is reproduced from the published findings. It shows the relationship between starvation and fledging in Magellanic penguin chicks for the 28 years between 1983 and 2010.

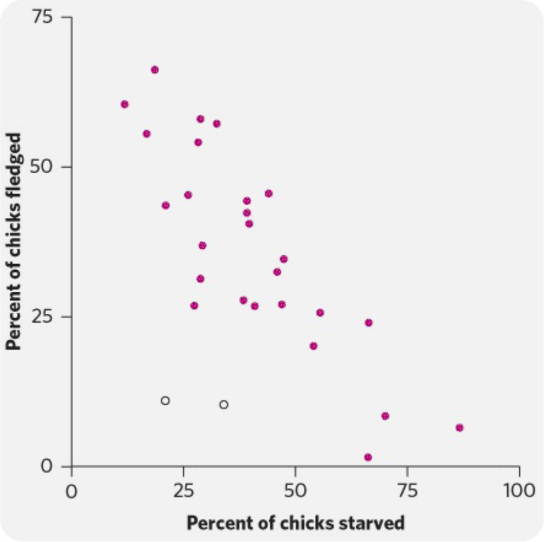


Figure 3.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 3.3 Scatterplot of the relationship between starvation and fledging in Magellanic penguin chicks between 1983 and 2010. Data for two years with extremely unusual weather are shown as open symbols.

Wayne Lynch/All Canada
/Shutterstock.com

The scatterplot shows a clear negative, linear pattern: Not surprisingly, years in which a higher percent of chicks died of starvation had a lower percent of chicks successfully fledging. However, two points (displayed with open symbols) stand noticeably apart from the overall pattern. These outliers of the relationship correspond to the years 1991

> and 1999, when torrential rain killed more than 40% of the chicks each year.
>
> The climate in Punta Tombo is arid, with low annual precipitation, so the years 1991 and 1999 were very unusual in this respect. For this reason, the researchers decided not to include these two data points when computing numerical summaries of the relationship between starvation and fledging.

Of course, not all relationships have a simple form and a clear direction that we can describe as positive association or negative association. Exercise 3.4 gives an example that does not have a single direction.

APPLY YOUR KNOWLEDGE

3.4 Enzyme activity and temperature. Enzymatic activity is known to be affected by temperature. A study examined the activity rate (in micromoles per second, μmol/s) of the digestive enzyme acid phosphatase in vitro at varying temperatures (measured in kelvins, K). The findings are displayed in the following table.[6] (For reference, the range of 253 to 298 K corresponds to roughly 25 to 80 degrees Celsius, or 77 to 176 degrees Fahrenheit.)

Temperature (K)	Rate (μmol/s) $(\mu\text{mol}/\text{s})$	Temperature (K)	Rate (μmol/s) $(\mu\text{mol}/\text{s})$	Temperature (K)	Rate (μmol/s) $(\mu\text{mol}/\text{s})$
298	0.04	318	0.34	338	1.05
298	0.05	318	0.34	338	1.10
298	0.05	318	0.35	338	1.06
303	0.08	323	0.47	343	0.96
303	0.08	323	0.48	343	0.98
303	0.08	323	0.48	343	0.99
308	0.11	328	0.78	348	0.72
308	0.13	328	0.78	348	0.73
308	0.11	328	0.79	348	0.75

313	0.18	333	0.97	353	0.54
313	0.20	333	0.98	353	0.58
313	0.19	333	1.03	353	0.62

a. Make a scatterplot. (Which is the explanatory variable?)
b. Describe the form of the relationship. It is not linear. Explain why the form of the relationship makes sense.
c. It does not make sense to describe the variables as strictly positively associated or negatively associated. Why?
d. s the relationship strong or weak? Explain your answer.

3.5 Chickadee alarm calls. The black-capped chickadee (*Poecile atricapilla*) is a small songbird commonly found in the northern United States and Canada. Chickadees often live in cooperative flocks, using a complex language to communicate about food sources and predator threats. In an experiment, researchers recorded chickadee vocalizations in an aviary when the birds were presented with predators of various sizes. The following data represent the average number of D notes per chickadee warning call for each type of predator, along with the predator wingspan (in centimeters):[7]

Predator	**Predator wingspan (cm)**	**Number of D notes per call**	**Predator**	**Predator wingspan (cm)**	**Number of D notes per call**
Pygmy owl	31.2	3.96	Gyrfalcon	115.1	2.25
Saw-whet owl	38.8	4.09	Peregrine falcon	120.0	2.80
Kestrel	57.6	2.76	Red-tailed hawk	120.0	2.56
Merlin	60.6	3.04	Great horned owl	120.4	2.46
Cooper’s hawk	80.6	3.18	Great gray owl	132.2	2.06
Short-eared owl	89.2	2.28	Rough-legged hawk	138.0	1.36
Prairie falcon	109.9	2.20			

a. Plot the number of D notes per call (response variable) against the predator wingspan (explanatory variable).
b. Describe the form, direction, and strength of the relationship between the number of D notes per call and the predator wingspan. What does the relationship suggest about chickadee warning calls?

ADDING CATEGORICAL VARIABLES TO SCATTERPLOTS

It is also possible to plot two or more relationships on the same scatterplot by using a different color or a different symbol for each.

CATEGORICAL VARIABLES IN SCATTERPLOTS

To add a categorical variable to a scatterplot, use a different plot color or symbol for each category.

Doing so allows us to examine visually the effect of a third variable on the relationship between x and y, a starting point to a **multivariate** analysis. First, describe the relationship between x and y separately for each category (that is, for each color or symbol). Then compare and contrast these relationships.[i]

EXAMPLE 3.7 The cost of reproduction in male fruit flies

Longevity in male fruit flies is positively associated with adult size. But do other factors, such as sexual activity, also matter? The cost of reproduction is well documented for the females of the species. A study looks at the association between longevity and adult size in male fruit flies kept under one of two conditions. One group is kept with sexually active females over their life span. The other group is cared for in the same way but kept with females that are not sexually active.[8]

Hermann Eisenbeiss/Science Source

Figure 3.4 shows the scatterplot of longevity (response variable) versus adult thorax length (indicative of body size, explanatory variable) for both conditions. The conditions are coded so that individual fruit flies from the sexually active group are represented by triangles, while the sexually inactive fruit flies from the second group are represented by circles. This coding introduces a third variable into the scatterplot. “Condition” is a categorical variable that has two values, identified by the two different plotting symbols.

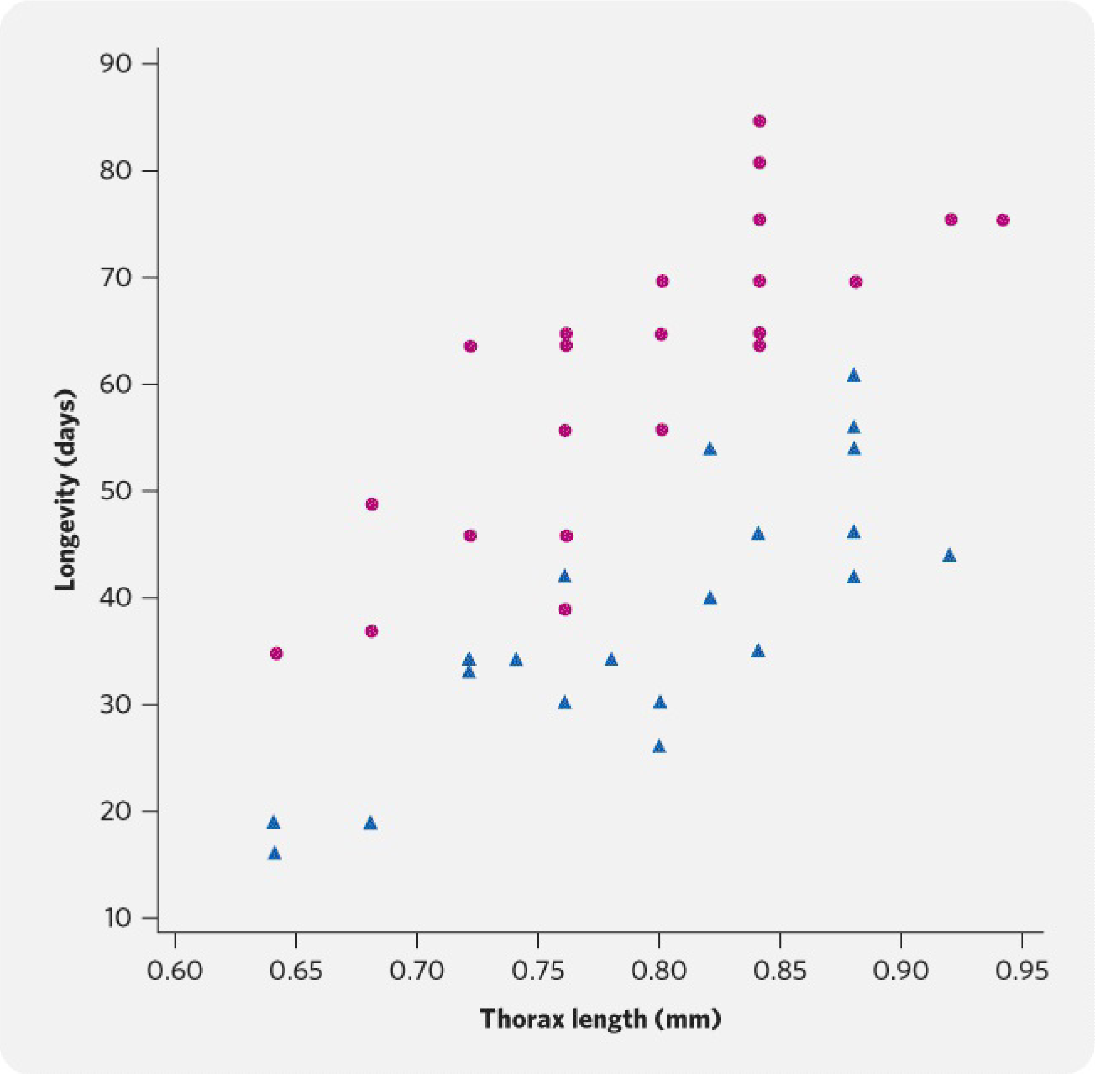


Figure 3.4

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 3.4 Cost of reproduction: scatterplot of thorax length (in millimeters) and longevity (in days) for sexually active (triangles) or inactive (circles) male fruit flies.

The scatterplot is quite clear. Both groups show a positive linear relationship between thorax length and longevity, as expected. However, the sexually active fruit flies have, on the whole, a lower longevity. That is, fruit flies of a given size tend to die sooner in the sexually active group than in the inactive group. In fact, only one male in the sexually active group outlived a male in the inactive group with similar thorax length.

APPLY YOUR KNOWLEDGE

3.6 Night watch. Sleeping in an unfamiliar environment can create a sleep disturbance called the first night effect. A sort of watchful sleep may occur if only one brain hemisphere truly exhibits sleep patterns. Researchers investigated asymmetric brain patterns in 11 healthy right-handed subjects on their first and second nights of sleep in the lab. The subjects' sleep latency (in minutes) was recorded each night and a noninvasive brain imaging technique was used to compute an index measuring the amount of asymmetric activity between the two brain hemispheres during deep sleep (a value of zero indicates no asymmetry). Here are the findings.[9]

Night 1		**Night 2**	
Sleep latency	Asymmetry index	Sleep latency	Asymmetry index
45	−0.107	17	0.047
12	−0.013	5	−0.021
26	−0.074	9	0.001
20	0.002	1	0.004
6	0.012	19	0.006
39	−0.055	17	0.015
30	0.019	3	0.020
22	−0.050	5	0.027
17	−0.045	15	−0.037
8	−0.009	4	0.009
9	0.001	5	0.014

a. Taking a long time to fall asleep could indicate a greater level of sleep disturbance, so sleep latency should be the explanatory variable. Display the data for both nights in the same scatterplot using different symbols for each night. Alternatively, you can plot the data in two side-by-side scatterplots that use the same x x and y y scales to make both patterns truly comparable.

b. Describe the relationship between sleep latency and brain asymmetry for each night, and then compare the patterns for both nights. What can you conclude about the brain's involvement in the first night effect?

Burger/Phanie/AGE Fotostock

3.7 Do heavier people burn more energy? Metabolic rate—the rate at which the body consumes energy—is important in studies of weight gain, dieting, and exercise. Here are data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight after taking out all fat. Metabolic rate is measured in kilocalories (Cal) burned per 24 hours (the same calories used to describe the energy content of foods). Researchers believe that lean body mass has an important influence on metabolic rate.

Subject	Sex	Mass (kg)	Rate (Cal)
1	M	62.0	1792
2	M	62.9	1666
3	F	36.1	995
4	F	54.6	1425
5	F	48.5	1396
6	F	42.0	1418
7	M	47.4	1362
11	F	40.3	1189
12	F	33.1	913
13	M	51.9	1460
14	F	42.4	1124
15	F	34.5	1052
16	F	51.1	1347
17	F	41.2	1204

8	F	50.6	1502	18	M	51.9	1867
9	F	42.0	1256	19	M	46.9	1439
10	M	48.7	1614				

a. Make a scatterplot of the data for the female subjects. Which is the explanatory variable? Explain why the subject number is not part of the scatterplot.

b. Is the association between these variables positive or negative? What is the form of the relationship? How strong is the relationship?

c. Now add the data for the male subjects to your graph, using a different color or a different plotting symbol. Does the pattern of relationship that you observed for women hold for men as well? How do the male subjects as a group differ from the female subjects as a group?

MEASURING LINEAR ASSOCIATION: CORRELATION

STATS IN YOUR WORLD

BODY MASS INDEX AND BODY FAT

Obesity and, more specifically, body fat have been linked to cardiovascular disease, strokes, and type 2 diabetes. Methods that directly measure body fat content, such as underwater weighing, dual-energy X-ray absorptiometry, or computerized tomography (CT) are costly and require professional intervention. By comparison, the body mass index (BMI) is calculated only from a person's height and weight. It has been shown to be strongly correlated with the results of the direct methods.

A scatterplot displays the direction, form, and strength of the relationship between two quantitative variables. Linear (straight-line) relations are particularly important because a straight line is a simple pattern that is quite common. A linear relation is strong if the points lie close to a straight line, and weak if they are widely scattered about a line. Like all qualitative judgments, however, statements such as "weak" and "strong" are open to individual interpretation. Therefore, we follow our strategy for data analysis by using a numerical measure to supplement the graph. *Correlation* is the measure we use.

CORRELATION

The **correlation** measures the direction and strength of the linear relationship between two quantitative variables. Correlation is usually written as r.r.

Suppose that we have data on variables xx and yy for nn individuals. The values for the first individual are x1x_1 and y1,y_1, the values for the second individual are x2x_2 and y2,y_2, and so on. The means and standard deviations of the two variables are x¯$\bar{x}$ and sxs_x for the xx-values, and y-$\bar{y}$ and sys_y for the yy-values. The correlation rr between xx and yy is r=1n−1∑(xi−x¯sx)(yi−y¯sy)

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

As always, the summation sign ∑ means “add these terms for all the individuals.” Computing rr by hand is a lengthy process. The formula for r r helps you to see what correlation is, and the next exercise asks you to calculate a correlation step by step from the definition to solidify its meaning. In practice, you should always use software or a calculator that finds rr from keyed-in values of two variables xx and y.$y.$ Note that some software programs refer to rr as **Pearson’s correlation** coefficient, or Pearson’s r.$r.$

The formula for rr begins by standardizing the observations. Suppose, for example, that xx is height in centimeters and yy is weight in kilograms and that we have height and weight measurements for nn people. Then x¯ $\bar{x}$ and s_x are the mean and the standard deviation of the nn heights, both in centimeters. The value

xi-x¯sx

$$\frac{x_i-\bar{x}}{s_x}$$

is the standardized height of the iith person. The standardized height says how many standard deviations above or below the mean a person’s height lies. Standardized values have no units—in this example, they are no longer measured in centimeters. Standardize the weights as well. Then, for each person, multiply the standardized height and the standardized weight. The correlation rr is, roughly speaking, an average of these products for the nn people.

The formula for correlation helps us see that rr is positive when there is a positive association between the variables. Height and weight, for example, have a positive association. People who are above average in height tend to also be above average in weight. For these people, both the standardized height and the standardized weight tend to be positive. People who are below average in height tend to also have below-average weight. For them, both standardized height and standardized weight tend to be negative. Overall, the products in the formula for rr are mostly positive, so

r is positive. In the same way, we can see that r is negative when the association between x and y is negative. More detailed study of the formula gives more detailed properties of r. Here is what you need to know to interpret correlation:

1. *Correlation makes no distinction between explanatory and response variables.* It makes no difference which variable you call x and which variable you call y when calculating the correlation.
2. Because r uses the standardized values of the observations, r *does not change when we change the units of measurement of x, y, or both.* Measuring height in inches rather than centimeters and weight in pounds rather than kilograms does not change the correlation between height and weight. The correlation r itself has no unit of measurement; it is just a number.
3. *In a linear association, a positive r indicates a positive association between the variables, and a negative r indicates a negative association.*
4. *The correlation r is always a number between −1 and 1.* Values of r near 0 indicate a very weak linear relationship. The strength of the linear relationship increases as r moves away from 0 toward either -1 or 1. Values of r close to -1 or 1 indicate that the points in a scatterplot lie close to a straight line. The extreme values $r=-1$ and $r=1$ occur only in the case of a perfect linear relationship, when the points lie exactly along a straight line.

EXAMPLE 3.8 From scatterplot to correlation

The scatterplots in Figure 3.5 illustrate how values of r closer to 1 or -1 correspond to stronger linear relationships. To make the meaning of r clearer, the standard deviations of both variables in these plots are equal, and the horizontal and vertical scales are the same. In general, it is not so easy to guess the value of r from the appearance of a scatterplot. Remember that changing the plotting scales in a scatterplot may mislead our eyes, but it does not change the correlation.

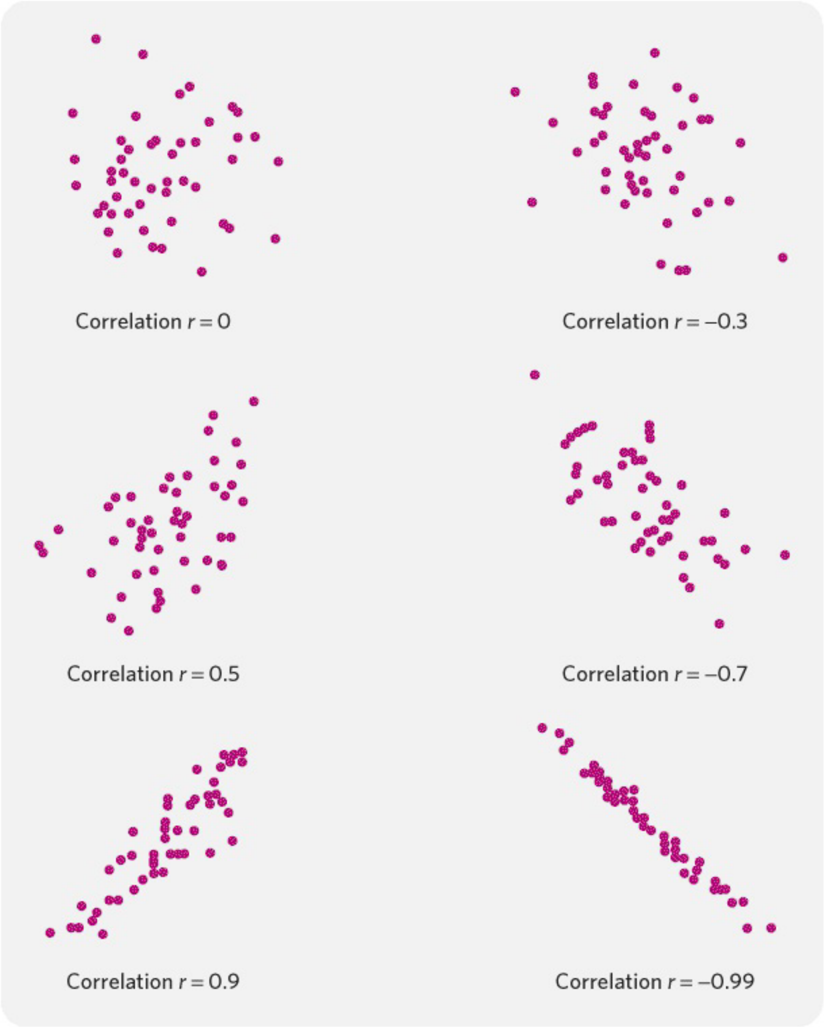


Figure 3.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 3.5 How correlation measures the strength of a linear relationship, for Example 3.8. Patterns closer to a straight line have correlations closer to 1

> 1 or −1.

> $$-1.$$
>
> The real data we have examined also illustrate how correlation measures the strength and direction of linear relationships. Figure 3.1 shows a very strong positive linear relationship between manatee deaths and powerboat registrations in Florida. The correlation in this case is r=0.945.$r = 0.945.$ Figure 3.2 shows a weaker but still fairly strong negative linear relationship between fat gain and nonexercise activity when overeating. Here the correlation is r=-0.779.$r = -0.779.$ Figure 3.4 shows a reasonably strong positive association between thorax length and longevity in the group of sexually active fruit flies (displayed as triangles). The correlation in this study is r=0.806.$r = 0.806.$ Notice how the strength of the association depends on the absolute value of rr (its numerical value irrespective of its sign).

Describing the relationship between two variables is a more complex task than describing the distribution of one variable. Here are some more facts about correlation and some cautions to keep in mind when you use r. $r.$

1. *Correlation requires that both variables be quantitative, so that it makes sense to perform the arithmetic indicated by the formula for r.*$r.$ For example, we cannot calculate a correlation between the amount of fat in the diets of a group of people and their ethnicity, because ethnicity is a categorical variable.
2. Correlation measures the strength of only the linear

relationship between two variables. *Correlation does not describe curved relationships between variables, no matter how strong they are.* Exercise 3.9 illustrates this important fact.

3. *Like the mean and the standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations.* In Example 3.6 we saw that researchers had identified two outliers (shown as open symbols in Figure 3.3) representing years with exceptional flood events. The correlation between starvation and fledging is r=-0.68$r = -0.68$ when the two outliers are excluded, but it is a much weaker −0.47 when they are included. Use rr with caution when the scatterplot includes any outliers.
4. *Correlation calculated from averaged data is typically much stronger than correlation calculated from the raw individual data points* because averaged values mask some of the individual-to-individual variations that would otherwise appear in the scatterplot. Exercise 3.8 addresses this fact.
5. *Correlation is not a complete summary of two-variable data,* even after you have established that the relationship between the variables is linear. You should give the means and standard deviations of both xx and yy along with the correlation and sample size. Of course, these numerical summaries do not point out outliers or **clusters** in the scatterplot. Numerical summaries complement plots of data, but they don't replace them.

Lastly, we may come across data that have been collected over *time* instead of over a number of individuals. How do we handle these data?

In Table 3.1, manatee deaths and powerboats registered are listed for each year between 1977 and 2016. Because we wanted to understand how powerboats are associated with manatee deaths, we created a scatterplot of the relationship between the number of powerboats registered and the number of manatee deaths from powerboat collisions in any given year. We then analyzed this pattern regardless of time. The data in Table 3.1 could also be used to create a timeplot and study the trends over time for each variable separately. This would give us a historical perspective not available with the scatterplot.

It is possible to use time as the explanatory variable in a correlation analysis, though time in itself does not explain or cause the changes (other factors do). Exercise 3.8 provides such an example. All approaches have some merit. *What matters most is to carefully examine the question being investigated and to match the proper analytic method to the question.*

APPLY YOUR KNOWLEDGE

3.8 Breakup distress. College freshmen involved in a dating relationship agreed to participate in a nine-month-long study. Every two weeks, participants answered a detailed questionnaire about their relationship and their emotional well-being. Over the course of the study, 26 of the participants experienced a breakup. The average distress of these 26 participants at the time of breakup (0weeks) and over the following 10 weeks was assessed from their questionnaire answers. Here are the findings reported in a scientific publication:[10]

Time since breakup (weeks)	0	2	6	10
Average distress (unitless score)	3.3	3.1	2.9	2.6

a. Make a scatterplot. Which is the explanatory variable? Describe the relationship between average distress and time since breakup.

b. As a pedagogical exercise, find the correlation rr step by step. You may wish to round off to two decimal places in each step. First, obtain the mean and standard deviation of each variable. Second, compute the four standardized values for each variable. Finally, use the formula for r.r. Use your calculator or software to obtain rr and check that you get the same result as in your hand-calculations, up to roundoff error.

c. How strong is the correlation between average distress and time since breakup? Explain how the value for rr matches your description from the scatterplot. If we had measured time since breakup in days rather than weeks, how would that difference affect r?r?

d. Each value of yy in the data table represents the average emotional distress of 26 individuals at a specific time after breakup. If the published findings had included the distress values of all 26 participants rather than their average, would the corresponding scatterplot have shown more or less scatter? How would this affect the value of the correlation r?r? Explain your reasoning.

3.9 Enzyme activity and temperature, continued. In Exercise 3.4 you made a scatterplot of acid phosphatase activity rate at different temperatures and described the strength of the relationship. If you calculate the correlation based on the data provided in that exercise, you find r=0.81$r = 0.81$ (check for yourself). Explain why the numerical value of the correlation is only moderately strong, even though the scatterplot indicates that the effect of temperature on enzymatic activity is clear and very reliable.

3.10 Chickadee alarm calls, continued. Exercise 3.5 gives data on the number of D notes per chickadee warning call for predators with various wingspans.

a. Use your calculator to find the correlation between the number of D notes per call and wingspan.
b. The researchers also recorded chickadee calls in response to the introduction of a bobwhite, a harmless quail. The bobwhite has a wingspan of 48.1 cm, and the chickadees responded with an average of 1.74 D notes per call. Make a scatterplot of the data with this additional point. Find the new correlation when the data point for the bobwhite is included.
c. Measuring the chickadee response to the introduction of a bobwhite was meant as a contrast to the response to various predators. It is not surprising, then, that the bobwhite data point falls outside the pattern of responses to predators. Explain, in terms of what the correlation measures, why adding the bobwhite data point makes the correlation weaker (closer to 0).

CHAPTER 3 SUMMARY

- To study relationships between variables, we must measure the variables on the same group of individuals.
- If we think that a variable xx may explain or even cause changes in another variable y,y, we call xx an **explanatory variable** (also called an independent variable) and yy a **response variable** (also called a dependent variable).
- A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis (xx axis) and values of the other variable on the vertical axis (yy axis). Plot each individual's data as a point on the graph. Always plot the explanatory variable, if there is one, on the x x axis of a scatterplot.
- Plot points with different colors or symbols to see the effect of a categorical variable in a scatterplot.
- In examining a scatterplot, look first for an overall pattern showing the **form, direction,** and **strength** of the relationship, and then look for **outliers** or other deviations from this pattern.
- **Form: Linear relationships,** where the points show a straight-line pattern, are an important form of relationship between two variables. **Curved** relationships and **clusters** are other forms to watch for.
- **Direction:** If the relationship has a clear direction, we speak of either **positive association** (high values of the two variables tend to occur together) or **negative association** (high values of one variable tend to occur with low values of the other variable).
- **Strength:** The **strength** of a relationship is determined by how close the points in the scatterplot lie to a simple form such as a line.
- The **correlation** rr measures the strength and direction of the linear association between two quantitative variables xx and y.y. Although you can calculate a correlation for any scatterplot, rr measures only straight-

line relationships.

- Correlation indicates the direction of a linear relationship by its sign: r>0$r > 0$ for a positive association and r<0$r < 0$ for a negative association. Correlation always satisfies -1≤r≤1$-1 \leq r \leq 1$ and indicates the strength of a relationship by how close it is to -1-1 or 1.1. Perfect correlation, r= ±1,$r = \pm 1$, occurs only when the points on a scatterplot lie exactly on a straight line.
- Correlation ignores the distinction between explanatory and response variables. The value of rr is not affected by changes in the unit of measurement of either variable. Correlation is not resistant, so outliers can greatly change the value of r.r.

THIS CHAPTER IN CONTEXT

In this and the following two chapters we continue our study of exploratory data analysis with the goal of examining relationships *between* variables. Here we show how scatterplots are used to visualize patterns in the relationships between two quantitative variables. When this pattern is roughly linear, we can compute the quantity r, r, the correlation coefficient, to quantify the strength and direction of the observed relationship. Because the correlation coefficient loses its meaning when the form of a relationship is not linear, it is important to always plot and carefully examine your data first.

In Chapter 4 we will model an observed linear relationship by obtaining the equation of the straight line that best represents it. This model, the least-square regression line, can then be used to make predictions. We will also discuss the importance of remembering that an observed association does not imply a causal relationship.

In Chapter 23, we will expand correlation and regression into the area of statistical inference, asking whether the patterns we observe in our data hold for individuals that we have not observed. That is, we will ask whether the observed relationship could be explained by chance alone (like seeing a pattern in the clouds) or whether it represents an actual association in an entire population of individuals. Optional Chapter 28 will expand these ideas from the simple bivariate framework to more comprehensive multivariate analyses.

CHECK YOUR SKILLS

3.11 You have data for many individuals on their walking speed and their heart rate after a 10-minute walk. When you make a scatterplot, the explanatory variable on the x axis

a. is walking speed.
b. is heart rate.
c. doesn’t matter.

3.12 You have data for many individuals on their walking speed and their heart rate after a 10-minute walk. You expect to see

a. a positive association.
b. very little association.
c. a negative association.

3.13 Many species exhibit some degree of sexual dimorphism. Researchers examining the relationship between length and weight in a species of scorpion, *Centruroides vittatus*, should create a scatterplot of length and weight[11]

a. with different symbols for male and female scorpions.
b. for all scorpions regardless of sex, because sex is a categorical variable.
c. for only one sex of scorpion, because analyzing both sexes would introduce too much variability.

3.14 Figure 3.6 is a scatterplot of the age (in months) at which a child begins to talk and the child’s later score on a test of mental ability (Gesell Adaptive Score), for each of 21 children (2 sets of 2 children happen to have the same values).[12] To describe the relationship in Figure 3.6, you would say that

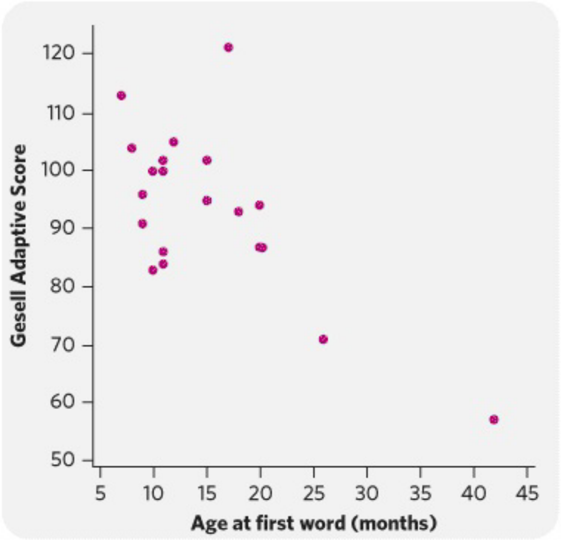


Figure 3.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 3.6 Scatterplot of Gesell Adaptive Score versus the age at first word for 21 children.

a. it is a positive curved association.
b. it is a negative linear association.
c. it shows no association.

3.15 In Figure 3.6, one child clearly falls outside the overall pattern of the scatterplot and, therefore, might be an outlier. This possible outlier is the child with

a. age 42 months and score 57.
b. age 26 months and score 71.
c. age 17 months and score 121.

3.16 What are all the values that a correlation rr can possibly take?

a. r≥0$r \geq 0$
b. 0≤r≤1$0 \leq r \leq 1$
c. -1≤r≤1$-1 \leq r \leq 1$

3.17 If mothers were always 2 years younger than the fathers of their children, the correlation between the ages of mothers and fathers would be

a. 1.
b. 0.5.

c. Can’t tell without seeing the data.

3.18 Because elderly people may have difficulty standing straight to have their height measured, a study looked at the relationship between overall height and height to the knee. Here are data (in centimeters) for five elderly men:

Knee height *x* (cm)	57.7	47.4	43.5	44.8	55.2
Overall height *y* (cm)	192.1	153.3	146.4	162.7	169.1

Using your calculator, what is the correlation between knee height and overall height?

a. r=0.88$r = 0.88$

b. r=0.09$r = 0.09$

c. r=0.77$r = 0.77$

3.19 In Exercise 3.18, both heights are measured in centimeters. Just for the fun of it, someone decides to measure knee height in millimeters and height in meters. The data in these units are

Knee height *x* (mm)	577	474	435	448	552
Overall height *y* (m)	1.921	1.533	1.464	1.627	1.691

The correlation for the data using these units would be

a. very close to the value calculated in Exercise 3.18.

b. exactly the same as in Exercise 3.18.

c. exactly 10 times smaller than in Exercise 3.18 (1 cm = 10 mm; 1 cm = 1/100 m).$(1\,\text{cm} = 10\,\text{mm}; 1\,\text{cm} = 1/100\,\text{m})$.

3.20 In Exercise 3.18 you calculated a correlation for knee height (x,x, in cm) and height (y,y, in cm). If, instead, you used height (in cm) as the xx variable and knee height (in cm) as the yy variable, the new correlation would

a. have the inverse value (1 over).

b. have the opposite value (a different sign).

c. remain the same.

CHAPTER 3 EXERCISES

3.21 Ecological approach to algal bloom control. Algal blooms can have negative effects on an ecosystem by dominating its phytoplankton communities. *Gonyostomum semen* is a nuisance alga infesting many parts of northern Europe. Could the overall biomass of *G. semen* be controlled by grazing zooplankton species? A research team examined the relationship between the net growth rate of *G. semen* and the number of *Daphnia magna* grazers introduced in test tubes. A net growth rate was computed by comparing the initial and final abundances of *G. semen* in the experiment, with a negative value indicating a decrease in abundance. Here are the findings:[13]

Number of grazers	1	2	3	4	5	6
Net growth rate	−1.9	−2.5	−2.2	−3.9	−4.1	−4.3

a. Make a scatterplot of number of grazers and net growth rate. Do you think that *D. magna* is an effective grazer of the *G. semen* alga?

b. Find the correlation r.r. How does it support your interpretation?

3.22 The amygdala and memory. The amygdala is a brain structure involved in the processing and memory of emotional reactions. In a research project, 10 subjects were shown emotional video clips. Their brains were then scanned with positron emission tomography (PET), and their memory of the video clips was assessed quantitatively. Here are the relative amygdala activity and the memory score for each subject:[14]

Relative activity	–0.417	–0.258	–0.234	–0.249	–0.156
Memory score	31	29	29	30	33
Relative activity	–0.031	0.120	0.240	0.342	0.654
Memory score	32	31	35	34	33

a. Make a scatterplot that shows the relationship between relative amygdala activity and memory score.

b. Describe the form of the relationship. Is there a linear pattern? How strong is the relationship? Compute the correlation coefficient.

c. Is the direction of the association positive or negative? Explain in simple language what this suggests about the involvement of the amygdala in the memory of emotional events.

3.23 More on algal bloom control. The study described in Exercise 3.21 also examined the grazing effect of another species, *Daphnia pulex*, on the abundance of the nuisance alga *G. semen* in the lab. Here are the findings:

Number of grazers	0	1	2	3	4	5	6
Net growth rate	−0.7	−0.4	−0.6	−0.3	−0.5	−1.0	−0.2

a. Make a scatterplot of number of grazers and net growth rate. Do you think that *D. pulex* is an effective grazer of the *G. semen* alga?

b. Find the correlation r.r. How does it support your interpretation?

c. The correlation between number of *D. magna* and net growth rate of *G. semen* in Exercise 3.21 was r=-0.93.$r = -0.93$. Contrast the findings for the two grazing species and state your conclusions about the potential effectiveness of *D. magna* and *D. pulex* as ecological methods to control *G. semen* algal blooms.

3.24 Night watch, continued. In Exercise 3.6 you plotted the relationship between sleep latency and brain asymmetry for healthy adults on their first and second nights sleeping in an unfamiliar setting. Obtain the correlation rr for each night. Then write a short paragraph describing the study findings, citing the correlation values to support your conclusion.

3.25 Global climate change, continued. Exercise 3.3 described a 2015 report by the Pew Research Center relating a country's per-capita CO_2 emissions to its citizens' level of concern over global climate change. The data from the report are displayed in Table 3.2.

TABLE 3.2 CO2CO_2 emissions (in ton per capita) and climate change concern i

Country	CO2 CO_2	Concern	Country	CO2 CO_2	Concern	Country	CO2 CO_2
Argentina	4.67	10.70	Ghana	0.41	10.73	Mexico	3.91
Australia	16.52	8.75	India	1.70	10.77	Nigeria	0.54

Brazil	2.23	11.42	Indonesia	2.31	9.21	Pakistan	0.93
Burkina Faso	0.11	11.22	Israel	8.95	8.66	Palestine	0.57
Canada	14.14	9.45	Italy	6.70	10.12	Peru	1.79
Chile	4.59	10.86	Japan	9.29	10.11	Philippines	0.86
China	6.71	9.11	Jordan	3.60	9.26	Poland	8.34
Ethiopia	0.08	9.49	Kenya	0.32	10.41	Russia	12.65
France	5.19	9.94	Lebanon	4.67	9.92	Senegal	0.59
Germany	8.92	9.49	Malaysia	7.85	9.86	South Africa	9.26

a. Create a scatterplot of the data and describe the relationship. If your country was part of this survey, locate it on the scatterplot.

b. How strong is this relationship? Obtain the correlation.

c. Refer to your answer to Exercise 3.3, part a. Would it be appropriate to conclude that citizens' level of concern over global climate change is directly affected by their country's CO_2 emissions? Explain your reasoning.

3.26 Grading oysters. Oysters are categorized for retail as small, medium, or large based on their volume. The grading process is slow and expensive when done by hand. A computer program estimates oyster volume based on the pixel area of two-dimensional (2D) images of the oysters. Table 3.3 shows the actual volume (in cm^3) and 2Dreconstruction (in thousands of pixels) for 30 oysters.[15]

TABLE 3.3 Actual volume and computer reconstruction of 30 oysters

Actual (cm^3)	2D (thousand pixels)	3D (million voxels)	Actual (cm^3)	2D (thousand pixels)	3D (million voxels)	Actual (cm^3)	2D (thousand pixels)	(m vo
13.04	47.907	5.136699	10.53	31.216	3.942783	10.95	37.156	4.7
11.71	41.458	4.795151	10.84	41.852	4.052638	7.97	29.070	3.0
17.42	60.891	6.453115	13.12	44.608	5.334558	7.34	24.590	2.7
7.23	29.949	2.895239	8.48	35.343	3.527926	13.21	48.082	4.9
10.03	41.616	3.672746	14.24	47.481	5.679636	7.83	32.118	3.1
15.59	48.070	5.728880	11.11	40.976	4.013992	11.38	45.112	4.4
9.94	34.717	3.987582	15.35	65.361	5.565995	11.22	37.020	4.5
7.53	27.230	2.678423	15.44	50.910	6.303198	9.25	39.333	3.4

12.73	52.712	5.481545	5.67	22.895	1.928109	13.75	51.351	5.
12.66	41.500	5.016762	8.26	34.804	3.450164	14.37	53.281	5.

a. We want to know whether the 2D imaging program is a good predictor of oyster volume, so 2D reconstruction is our explanatory variable. Make a scatterplot of 2D reconstruction and actual volume.

b. Describe the overall pattern of the relationship: its form, direction, and strength. If appropriate, give the correlation coefficient. Would you say that the 2D system is an accurate way to assess oyster volume?

3.27 More on grading oysters. Exercise 3.26 describes an automated grading system for oysters. Engineers were given the task of improving on the 2D reconstruction program. They designed a new program that estimates oyster volume using three-dimensional (3D) digital image processing. The results are displayed in Table 3.3.

a. Make a scatterplot of 3D volume reconstruction (in millions of volume pixels) and actual volume (in cm^3), using 3D reconstruction as the explanatory variable. Describe the overall pattern of the relationship and, if appropriate, give the correlation coefficient.

b. Compare your analysis for this 3D system with that for the 2D system from Exercise 3.26. Is the 3D reconstruction program an improvement over the 2D version? Explain your reasoning.

3.28 Global warming and beer quality. Hops are the flower clusters of a perennial herbaceous plant, *Humulus lupulus*, used in brewing to facilitate fermentation and add a bitter flavor to beer. Hops with a higher α-acid $\alpha - \text{acid}$ content provide more bitterness, and α-acid$\alpha - \text{acid}$ content is used as a criterion for hops quality. A study examined how climate change is affecting α-acid$\alpha - \text{acid}$ content in hops. Temperature records and data on hop production in the Czech Republic over the 1954–2006 period were collected. Table 3.4 shows the annual average daily temperature from April to August (in degrees Celsius) and the annual percent of α-acids $\alpha - \text{acids}$ in hops during that time period. The data are sorted by temperature, not by year.[16]

TABLE 3.4 **Annual average daily temperature and annual percent of αα-acids**

Temperature (° C)	Percent α-acids$\alpha - \text{acids}$	Temperature (° C)	Percent α-acids$\alpha - \text{acids}$	Temperature (° C)	Percent α-acids$\alpha - \text{acids}$
13.34	4.39	14.53	6.09	15.21	6.21
13.59	4.39	14.53	5.59	15.22	4.10
13.66	5.63	14.62	4.82	15.33	4.10
13.83	6.30	14.69	5.12	15.36	4.69
13.95	5.49	14.71	5.23	15.40	5.60
14.01	4.10	14.76	4.39	15.44	3.51
14.13	4.10	14.82	5.81	15.53	5.91
14.19	5.29	14.90	5.50	15.56	3.99
14.24	5.10	14.91	5.84	15.61	3.59
14.24	4.41	14.92	4.36	15.65	4.18
14.28	5.01	15.02	4.91	15.80	3.80
14.37	5.80	15.06	4.99	15.91	3.61
14.45	4.75	15.09	3.70	15.98	3.30
14.47	5.48	15.19	6.21	16.02	3.80

a. The researchers examined the pattern of change over time in annual percent of α-acids.$\alpha - \text{acids}.$ What type of graph do you think they used? Explain your reasoning.

b. Make a graph showing the relationship between average daily temperature and percent of α-acids$\alpha - \text{acids}$ in the same year. Which variable would you assign to be the explanatory variable, if any? Describe the relationship revealed by your graph.

a. If appropriate, calculate a numerical summary of the strength of the relationship between average daily temperature and percent of α-acids. $\alpha - \text{acids}.$ What are your conclusions about the impact of higher average daily temperature on hops quality?

3.29 Huntington’s disease. Huntington’s disease (HD) is a genetic neurodegenerative disorder. Patients with HD experience progressive brain degeneration, especially in the caudate nucleus. Researchers examined the association between the molecule *N*-acetylaspartate (NAA, a brain-specific metabolite) in the caudate nucleus and caudate atrophy in a sample of 10 patients with very early-stage HD. Here are the patients' caudate NAA concentrations (in international units, IU) and caudate volumes (in milliliters, ml):[17]

Subject	Volume	NAA	Subject	Volume	NAA
1	2.6	9.4	6	3.1	8.4
2	2.5	7.8	7	4.0	10.9
3	2.9	7.6	8	3.5	9.5
4	3.5	10.0	9	4.2	11.2
5	3.4	8.8	10	3.7	11.6

a. It is currently unknown whether caudate atrophy might precede or follow any change in NAA level. Therefore, the variables here are neither explanatory nor response variables. Display and describe the relationship between caudate volume and caudate NAA level.

b. How strong is the association between caudate volume and caudate NAA level? Give the value of the correlation.

3.30 Energy cost of running. If you run on a computerized treadmill at the gym and check the calories you have just burned, chances are that the machine's answer is based on the results of a 1963 study. In this study, researchers asked athletes to run on a treadmill at various speeds and inclines and assessed the athletes' energy expenditure (computed indirectly via oxygen consumption and individual body measurements). Figure 3.7 shows 25 measurements of energy expenditure measured under various conditions of speed (in km/h) and treadmill incline (in percent). Each incline is coded by a different color on the graph.[18]

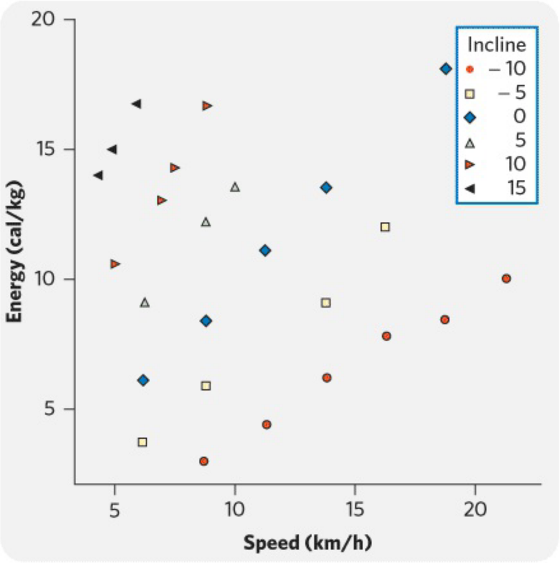


Figure 3.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 3.7 Scatterplot of energy expenditure against running speed, using different colors and symbols to represent various running inclines.

a. Describe the relationship between energy expenditure and speed (form, direction, and strength) separately for each incline.
b. How does the treadmill incline affect the relationship between energy expenditure and speed?

3.31 Neural mechanism of nociception. The roundworm *Caenorhabditis elegans* is a widely studied animal model, in part because of its small number of neurons and easily manipulated genome. Nociception is the neural perception of an actually or potentially harmful stimulus. In *C. elegans*, it evokes a self-preserving withdrawal behavior. However, repeated stimulation can result in reduced withdrawal response, or habituation. Researchers compared the withdrawal response to disturbing light stimuli in wild-type *C. elegans* and a mutant *C. elegans* line that exhibits a slower response of its nociceptive sensory neurons (PVD). Figure 3.8 shows the scatterplot of the percent of animals tested that exhibited a withdrawal reaction to a noxious stimulus consisting of varying numbers of consecutive light pulses. Failure to react indicates habituation. Circles represent wild-type *C. elegans*, and squares represent the mutant line.[19]

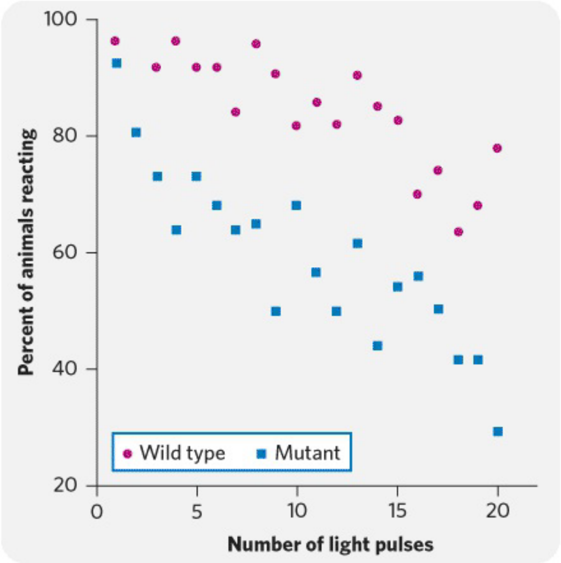


Figure 3.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 3.8 Scatterplot of the percent of animals reacting to a noxious luminous stimulus consisting of varying numbers of consecutive light pulses. Circles represent wild-type *C. elegans*, and squares represent mutant *C. elegans*.

a. What does the scatterplot show about the pattern of withdrawal responses in wild-type *C. elegans* for increasing numbers of light pulses? How does your answer fit in the context of habituation?
b. What does the scatterplot show about the difference between wild-type and mutant *C. elegans* lines? Explain why these results suggest an involvement of PVD sensory neurons in nociception and habituation.

3.32 Transforming data: counting seeds. Table 3.5 shows data on the mean number of seeds produced in a year by several common tree species and the mean weight (in milligrams) of the seeds produced. (Some species appear twice because their seeds were counted in two locations.) We might expect that trees with heavy seeds would produce fewer seeds, but what is the actual form of the relationship?[20]

TABLE 3.5 Count and weight of seeds produced by common tree species

Tree species	Seed count	Seed weight (mg)	Tree species	Seed count	Seed weight (mg)
Paper birch	27,239	0.6	American beech	463	247

Yellow birch	12,158	1.6	American beech	1,892	247
White spruce	7,202	2.0	Black oak	93	1,851
Engelmann spruce	3,671	3.3	Scarlet oak	525	1,930
Red spruce	5,051	3.4	Red oak	411	2,475
Tulip tree	13,509	9.1	Red oak	253	2,475
Ponderosa pine	2,667	37.7	Pignut hickory	40	3,423
White fir	5,196	40.0	White oak	184	3,669
Sugar maple	1,751	48.0	Chestnut oak	107	4,535
Sugar pine	1,159	216.0			

a. Make a scatterplot showing how the weight of tree seeds helps explain how many seeds the tree produces. Describe the form, direction, and strength of the relationship.

b. When dealing with sizes and counts, the logarithms of the original data are often a better choice of variable. Use your calculator or software to obtain the logarithms of both the seed weights and the seed counts in Table 3.5. Make a new scatterplot using these new variables. What are the form, direction, and strength of the relationship appearing in this new scatterplot? Most software programs provide the option of using logarithm scales for the axes of scatterplots, allowing you to skip the conversion to logarithms of the original data.

3.33 Transforming data: cancer. Some types of cancer are much more common than others. Although genetic and environmental factors contributing to cancer often make the news, cancer may also arise from random mutations during routine stem cell divisions over the course of a lifetime. Researchers examined the relationship between the total number of stem cell divisions in the lifetime of a given tissue and the lifetime risk of cancer in that tissue. Table 3.6 provides the data for 31 types of cancers for which this information is known in the U.S. population. Each risk value reflects the proportion of individuals in the population who get this particular cancer over their lifetime.[21]

TABLE 3.6 **Lifetime risk and number of stem cell divisions for 31 cancer types**

Cancer type	Lifetime risk	Stem cell divisions	Cancer type	Lifetime risk	S
AM leukemia	0.0041000	129,900,000,000	Lung (smokers)	0.0810000	9,
Basal cell	0.3000000	3,550,000,000,000	Medulloblastoma	0.0001100	
CL leukemia	0.0052000	129,900,000,000	Melanoma	0.0203000	763,
Colorectal	0.0480000	1,168,000,000,000	Osteosarcoma	0.0003500	
FAP colorectal	1.0000000	1,168,000,000,000	Arms osteosarcoma	0.0000400	
Lynch colorectal	0.5000000	1,168,000,000,000	Head osteosarcoma	0.0000302	
Duodenum	0.0003000	7,796,000,000	Legs osteosarcoma	0.0002200	
FAP duodenum	0.0350000	7,796,000,000	Pelvis osteosarcoma	0.0000300	
Esophageal	0.0019380	1,203,000,000	Ovarian germ cell	0.0004110	
Gallbladder	0.0028000	78,400,000	Pancreatic ductal	0.0135890	342,
Glioblastoma	0.0021900	270,000,000	Pancreatic islet	0.0001940	6,
Head and neck	0.0138000	31,860,000,000	Small intestine	0.0007000	292,
HPV head and neck	0.0793500	31,860,000,000	Testicular	0.0037000	3,
Hepatocellular	0.0071000	270,900,000,000	Thyroid follicular	0.0102600	
HCV hepatocellular	0.0710000	270,900,000,000	Thyroid medullary	0.0003240	
Lung (nonsmokers)	0.0045000	9,272,000,000			

a. Make a scatterplot showing lifetime cancer risk as a function of total number of stem cell divisions. Describe the form, direction, and strength of the relationship. Would it be appropriate to compute the correlation r between these two variables? Explain your reasoning.

b. When a variable spreads over several orders of magnitude, transforming the data using a base-10 logarithm function helps focus on these differences of magnitude. What are the minimum and maximum for each variable? Now obtain the base-10 log of each value in the table. What are the minimums and the maximums now?

c. Make a scatterplot showing the log of lifetime cancer risk as a function of the log of total number of stem cell divisions. Describe the form, direction, and strength of this log–log relationship. (Most software programs provide the option of using logarithm scales for the axes of scatterplots, allowing you to skip the conversion to logarithms of the original data.) How strong is the relationship between the log of cancer risk and the log of stem cell divisions? Obtain the value of r using the log-transformed data.

3.34 Merlins breeding. Often the percent of an animal species in the wild that survives to breed again is lower following a successful breeding season. This pattern is part of nature's self-regulation to keep population size stable. A study of merlins (small falcons) in northern Sweden observed the number of breeding pairs in an isolated area and the percent of males (banded for identification) who returned the next breeding season. Here are data for nine years:[22]

Breeding pairs	28	29	29	29	30	32	33	38	38
Percent return	82	83	70	61	69	58	43	50	47

Do the data support the theory that a smaller percent of birds survive following a successful breeding season? Use the four-step outline (page 58) in your answer.

3.35 Bushmeat. Bushmeat—the meat of wild animals—is widely traded in Africa, but its consumption threatens the survival of some animals in the wild. Bushmeat is often not the first choice of consumers: They eat bushmeat only when other sources of protein are in short supply. Researchers looked at declines in 41 species of mammals in nature reserves in Ghana and at catches of fish (the primary source of animal protein) in the same region. The data appear in Table 3.7.[23] Fish supply is measured in kilograms per person. The other variable is the percent change in the total "biomass" (weight in tons) for the 41 animal species in six nature reserves. Most of the yearly percent changes in wildlife mass are negative because most years saw fewer wild animals in West Africa. Follow the four-step outline (page 58) to examine whether the data support the idea that more animals are killed for bushmeat when the fish supply is low.

TABLE 3.7 **Fish supply and wildlife decline in West Africa**

Year	Fish supply (kg per person)	Biomass change (percent)	Year	Fish supply (kg per person)	Biomass change (percent)
1971	34.7	2.9	1985	21.3	−5.5
1972	39.3	3.1	1986	24.3	−0.7
1973	32.4	−1.2	1987	27.4	−5.1
1974	31.8	−1.1	1988	24.5	−7.1
1975	32.8	−3.3	1989	25.2	−4.2
1976	38.4	3.7	1990	25.9	0.9
1977	33.2	1.9	1991	23.0	−6.1
1978	29.7	−0.3	1992	27.1	−4.1
1979	25.0	−5.9	1993	23.4	−4.8
1980	21.8	−7.9	1994	18.9	−11.3
1981	20.8	−5.5	1995	19.6	−9.3
1982	19.7	−7.2	1996	25.3	−10.7
1983	20.8	−4.1	1997	22.0	−1.8
1984	21.1	−8.6	1998	21.0	−7.4

3.36 Does social rejection hurt? We often describe our emotional reaction to social rejection as “pain.” A clever study asked whether social rejection causes activity in areas of the brain that are known to be activated by physical pain. If it does, we really do experience social and physical pain in similar ways. Subjects were first included and then deliberately excluded from a social activity while changes in brain activity were measured. After each activity, the subjects filled out questionnaires that assessed how excluded they felt. Here are data for 13 subjects:[24]

Subject	Social distress	Change in brain activity	Subject	Social distress	Change in brain activity
1	1.26	−0.055	8	2.18	0.025
2	1.85	−0.040	9	2.58	0.027
3	1.10	−0.026	10	2.75	0.033

4	2.50	−0.017	11	2.75	0.064
5	2.17	−0.017	12	3.33	0.077
6	2.67	0.017	13	3.65	0.124
7	2.01	0.021			

The explanatory variable is “social distress,” measured by each subject’s questionnaire score after exclusion relative to the score after inclusion. (Thus, values greater than 1 show the degree of distress caused by exclusion.) The response variable is change in activity in a region of the brain that is activated by physical pain. Discuss what the data show. Follow the four-step process (page 58) in your discussion.

3.37 The toucan’s beak. The toco toucan, *Ramphastos toco*, is the largest member of the toucan family and possesses the largest beak relative to body size of all birds. This exaggerated feature has been interpretated in various ways, such as being a refined adaptation for feeding. However, the large surface area may also be an important mechanism for radiating heat (and hence cooling the bird) as outdoor temperature increases. Here are data for beak heat loss, as a percent of total body heat loss, at various temperatures in degrees Celsius:[25]

Temperature (°C)	15	16	17	18	19	20	21	22
Percent heat loss from beak	32	34	35	33	37	46	55	51
Temperature (°C)	23	24	25	26	27	28	29	30
Percent heat loss from beak	43	52	45	53	58	60	62	62

Investigate the relationship between outdoor temperature and beak heat loss as a percent of total body heat loss.

3.38 Correlation is not resistant. Go to the *Correlation and Regression* applet. Click on the scatterplot to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation about 0.9).

a. Add 1 point at the upper right that is in line with the first 10 points. How does the correlation change?

b. Drag this last point down until it is opposite the group of 10 points.

How small can you make the correlation? Can you make the correlation negative? You should see that a single outlier can greatly strengthen or weaken a correlation. For this reason, you should always plot your data to check for outlying points.

3.39 Match the correlation. In this exercise you will use the *Correlation and Regression* applet to make scatterplots with 10 points that have correlation close to 0.7. The lesson here is that many patterns can have the same correlation. Always plot your data before you trust a correlation.

a. Stop after adding the first 2 points. What is the value of the correlation? Why does it have this value?
b. Create a lower-left to upper-right pattern of 10 points with correlation about r=0.7.$r = 0.7$. (You can drag points up or down to adjust rr after you have 10 points.) Make a rough sketch of your scatterplot.
c. Make another scatterplot with 9 points in a vertical stack at the left of the plot. Add 1 point far to the right, and then move it until the correlation is close to 0.7. Make a rough sketch of your scatterplot.
d. Make yet another scatterplot with 10 points in a curved pattern that starts at the lower left, rises to the right, and then falls again at the far right. Adjust the points up or down until you have a quite smooth curve with correlation close to 0.7. Make a rough sketch of this scatterplot.

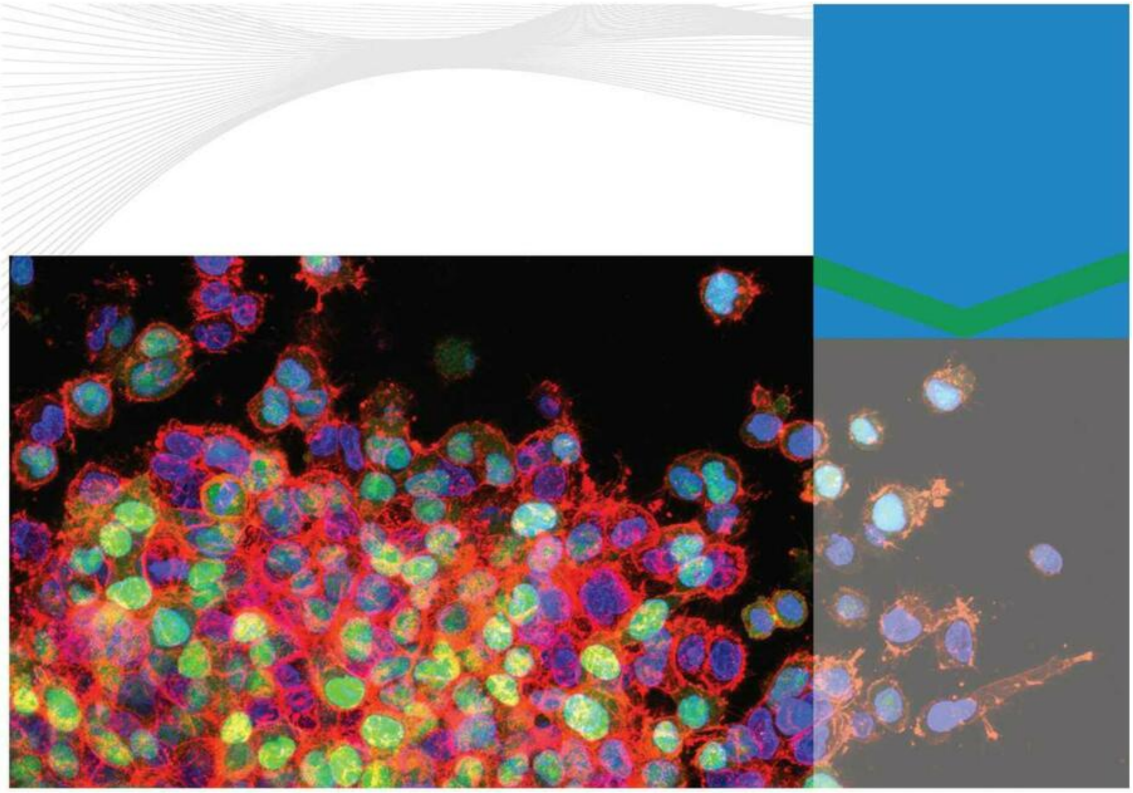
Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Short answers are not available for these exercises to enable their use for class work or assignments.*

3.40 Everglades. Continue your work from Exercise 1.45 (page 36) with the data file *Large.Everglades.*

a. Make a scatterplot matrix displaying the relationship for each combination of 5 variables. There should be 10 different graphs (20 graphs if each variable is used on the xx axis and then on the yy axis for each combination). Which relationships, if any, are linear?

b. The Shark River ends in the Gulf of Mexico. When the tide comes in, ocean water flows into the river; when the tide goes out, water flows from the river into the ocean. The velocity and discharge variables take positive values when the tide is out, and negative values when the tide is in. Describe again the scatterplot matrix you created in part a, this time taking into account the tides' impact (that is, describe the tide in and the tide out separately).

CHAPTER 4 Regression



National Cancer Institute/Winship Cancer Institute of Emory University

IN THIS CHAPTER WE COVER...

- The Least-Squares Regression Line
- Facts About Least-Squares Regression
- Outliers and Influential Observations
- Working with Logarithm Transformations
- Cautions About Correlation and Regression
- Association Does not Imply Causation

Linear (straight-line) relationships between two quantitative variables are easy to understand and quite common. In Chapter 3 we found linear relationships in settings as varied as manatee deaths, weight gain, and concern over climate change. Correlation measures the direction and strength of these relationships. When a scatterplot shows a linear relationship, we may want to summarizes the overall pattern by drawing a line on the scatterplot.

REVIEW OF STRAIGHT LINES

Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y=a+bx$$

$y = a + bx$

In this equation, b is the **slope,** the amount by which y changes when x increases by one unit. The number a is the **intercept,** the value of y when $x=0$.

THE LEAST-SQUARES REGRESSION LINE

A **regression line** is a straight line that summarizes the linear relationship between two variables, but only in a specific setting: When one of the variables is thought to help explain or predict the other. That is, regression describes a relationship between an explanatory variable (x) and a response variable (y).

A regression line is a mathematical **model.** It represents the underlying pattern in the data and provides an equation showing how the response variable y changes, *on average*, as the explanatory variable x changes. This equation may be used to predict the *expected* value of y for a given value of x.

In Example 3.5 (page 72), we examined the relationship between fat gain (in kilograms, kg) and change in nonexercise activity (NEA in kilocalories, or Calories, Cal) in young adult volunteers who had overeaten for eight weeks. We found that those with larger NEA increases tended to gain less fat. To add to this conclusion, we must study regression lines in more detail, starting with the basic concepts.

EXAMPLE 4.1 Using a regression line

Figure 4.1 repeats Figure 3.2 (page 73), with the addition of a regression line for predicting fat gain from NEA change. What does this line mean? Suppose that an individual’s NEA increases by 400 Cal when she overeats. From 400 Cal on the x axis, go up to the regression line and then over to the y axis. The graph shows that the fat gain predicted by the linear model is a bit more than 2 kg.

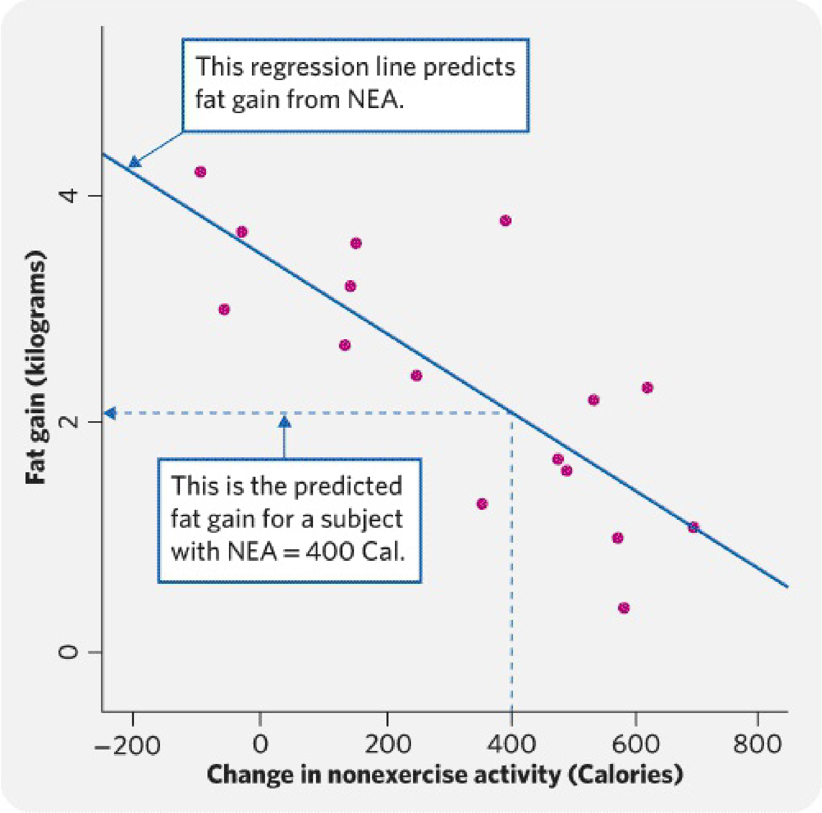


Figure 4.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 4.1 Fat gain after 8 weeks of overeating, plotted against increase in nonexercise activity over the same period.

Any straight line describing the NEA data has the form

$$\text{fat gain} = a + (b \times \text{NEA change})$$

$$\text{fat gain} = a + (b \times \text{NEA change})$$

where “fat gain” is y and “NEA change” is x. The line in Figure 4.1 is the regression line with the equation

$$\text{fat gain} = 3.505 - (0.00344 \times \text{NEA change})$$

$$\text{fat gain} = 3.505 - (0.00344 \times \text{NEA change})$$

This equation makes it easy to predict fat gain. If a person’s NEA increases by 400 Cal when she overeats, substitute $x = 400$ in the equation. The fat gain predicted by this model is

$$\text{fat gain} = 3.505 - (0.00344 \times 400) = 2.13\ \text{kg}$$

$$\text{fat gain} = 3.505 - (0.00344 \times 400) = 2.13\ \text{kg}$$

Be sure you understand what this predicted fat gain of 2.13 kg means. It is our best guess, based on the linear model, for one individual whose NEA would increase by 400 Cal when overeating. However, different individuals may actually gain more or less fat for the same NEA increase. The scatterplot shows this very clearly. The value 2.13 kg represents the estimated fat gain, *on average*, for all individuals who increase NEA by 400 Cal when overeating.

The regression equation itself can also provide useful information in context.

- The slope, $b = -0.00344$, tells us that fat gained decreases by 0.00344 kg, *on average*, for each added Calorie of NEA. The slope of a regression line is the expected *rate of change* in the response variable as the explanatory variable changes.
- The intercept, $a = 3.505$ kg, is the estimated fat gain if NEA does not change ($x = 0$) when a person overeats.

The slope of a regression line is an important numerical description of the relationship between the two variables and it always has a relevant meaning in the context of the problem. In contrast, *although we need the value of the intercept to complete the regression equation, this value is meaningful in context only when the explanatory variable can actually take values close to zero.* Consider, for instance, the relationship between time to finish a one-mile race (x) and pulse rate at the finish line (y). Because it is physiologically impossible to run one mile in zero time, the value of the y-intercept would have no useful interpretation in this particular context.

Notice that the slope $b = -0.00344$ in Example 4.1 is small. This does *not* mean that change in NEA has little effect on fat gain. *The size of the slope depends on the units in which we measure the two variables.* In this example, the slope is the change in fat gain in kilograms when NEA increases by 1 Cal. There are 1000 grams in a kilogram. If we measured fat gain in grams, the slope would be 1000 times larger, $b = 3.44$. *You can't determine how important a relationship is by looking at the size of the slope of the regression line.*

How do we find a regression line on a scatterplot that we can all agree on and that does not depend on an approximate guess? Because we use the line to predict yy from xx, the prediction errors we make are errors in yy, the vertical direction in the scatterplot. The vertical distances of the points from the line are called **residuals.** A good regression line makes the residuals as small as possible. The most common way to make the collection of vertical prediction errors “as small as possible” is the *least-squares* method.

LEAST-SQUARES REGRESSION LINE

The **least-squares regression line** of yy on xx is the line that makes the sum of the squared vertical distances of the data points from the line as small as possible.

One reason for the popularity of the least-squares regression line is that the problem of finding the line has a simple answer. The equation for the least-squares regression line can be expressed in terms of the means and standard deviations of the two variables and the correlation between them.

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

We have data on an explanatory variable xx and a response variable yy for nn individuals. From the data, we can calculate the means x¯ $\bar{x}$ and y¯ $\bar{y}$ and the standard deviations sxs_x and sy s_y of the two variables and their correlation rr. The least-squares regression line is the line

y^=a+bx

$$\hat{y} = a + bx$$

with slope

b=rsysx

$$b = r\frac{sy}{sx}$$

and intercept

a=y¯−bx¯

$$a = \bar{y} - b\bar{x}$$

We write y^$\hat{y}$ (read “y hat”) in the equation of the regression line to emphasize that the line is a mathematical model of the underlying linear pattern that gives a *predicted* (expected, or average) response y^$\hat{y}$ for any xx. Because of the scatter of points about the line, the predicted response will usually not be exactly the same as the actually *observed* response yy. The differences y-y^$y - \hat{y}$ are the residuals of the regression.

EXAMPLE 4.2 Finding the regression line

Example 4.1 gives the equation for the least-squares regression line between NEA change and fat gain. We can obtain this equation fairly quickly. From the raw data given in Example 3.5 (page 72), we find that the sample means and standard deviations for the two variables are

	NEA change	**Fat gain**
Mean	324.75	2.388
Standard deviation	257.66	1.139

The correlation between them is r=-0.779$r = -0.779$. This is all the information we need to find the slope and the intercept:

slope=rsysx=−0.7791.139257.66=−0.00344intercept=y¯−bx¯=2.388−(−0.00344)(324.75)=3.505

$$\begin{aligned} \text{slope} &= r\frac{s_y}{s_x} = -0.779\frac{1.139}{257.66} = -0.00344 \\ \text{intercept} &= \bar{y} - b\bar{x} = 2.388 - (-0.00344)(324.75) = 3.505 \end{aligned}$$

You should make sure to keep at least three significant digits, especially for the slope. *Rounding the slope too much can lead to substantial inaccuracies when trying to predict* yy *using the equation for the least-squares regression line.*

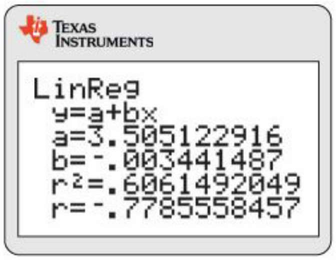
If all you have available is summary data such as means, standard deviations, and the correlation, you can use these formulas to find the equation for the least-squares regression line. In practice, however, any technology you use for statistical calculations will give you the least-squares line and related information.

STATS IN YOUR WORLD

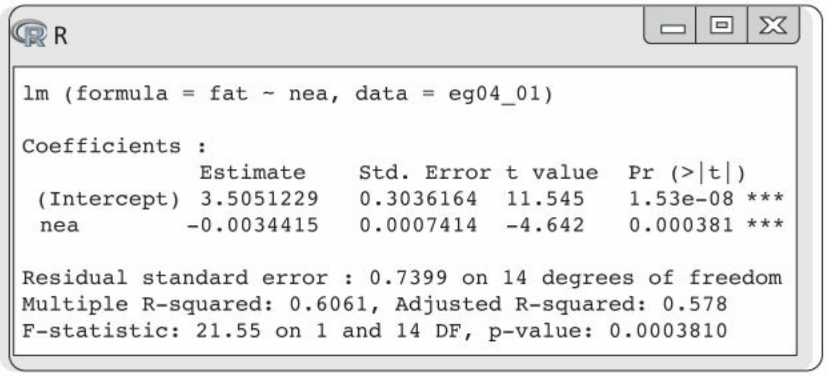
REGRESSION TOWARD THE MEAN

To “regress” means to go backward. Why are statistical methods for predicting a response from an explanatory variable called “regression”? Sir Francis Galton (1822–1911), who was the first to apply regression to biological and psychological data, looked at examples such as the heights of children versus the heights of their parents. He found that taller-than-average parents tended to have children who were also taller than average but not as tall as their parents. Galton called this fact “regression toward the mean,” and the label subsequently came to be applied to the statistical method.

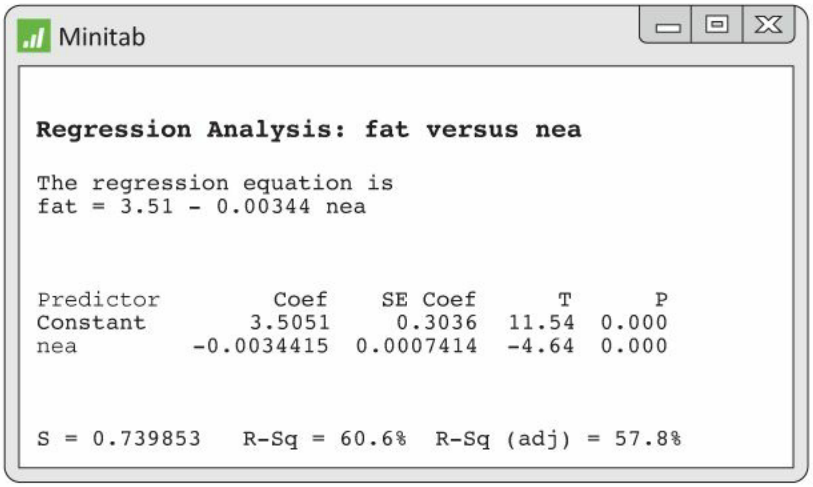
Figure 4.2 displays the regression output for the data of Example 4.2 from a graphing calculator, three statistical programs, and a spreadsheet program. Each output provides the intercept and slope of the least-squares line. The graphing calculator simply labels them aa and b,b, whereas the programs refer to them as “coefficients.” The yy intercept is listed as either “intercept” or “constant.” Note that the slope is often referred to by the name of the xx variable, here “nea.” Software also provides information that we do not yet need, although we will use much of it later.



```
LinReg
 y=a+bx
 a=3.505122916
 b=⁻.003441487
 r²=.6061492049
 r=⁻.7785558457
```

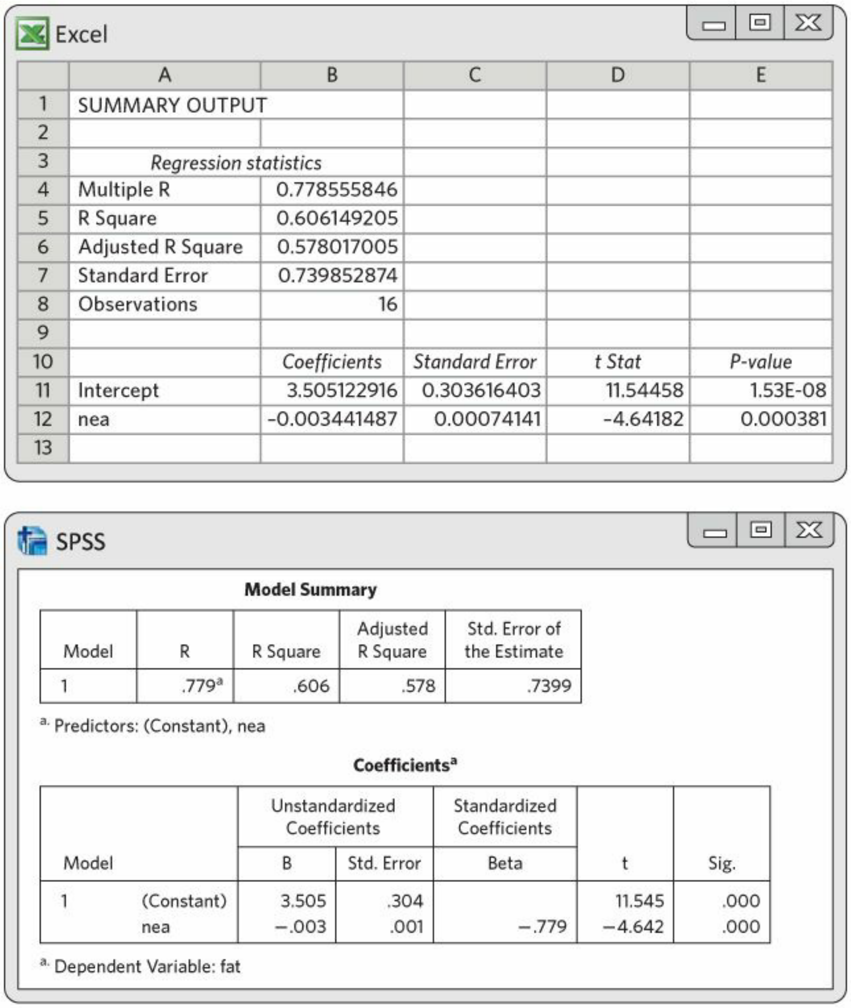


```
lm (formula = fat ~ nea, data = eg04_01)

Coefficients :
             Estimate    Std. Error t value   Pr (>|t|)
 (Intercept) 3.5051229   0.3036164  11.545    1.53e-08 ***
 nea        -0.0034415   0.0007414  -4.642    0.000381 ***

Residual standard error : 0.7399 on 14 degrees of freedom
Multiple R-squared: 0.6061, Adjusted R-squared: 0.578
F-statistic: 21.55 on 1 and 14 DF, p-value: 0.0003810
```



```
Regression Analysis: fat versus nea

The regression equation is
fat = 3.51 - 0.00344 nea


Predictor         Coef    SE Coef      T      P
Constant        3.5051     0.3036  11.54  0.000
nea         -0.0034415  0.0007414  -4.64  0.000


S = 0.739853   R-Sq = 60.6%   R-Sq (adj) = 57.8%
```

Figure 4.2 *(continued)*
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



	A	B	C	D	E
1	SUMMARY OUTPUT				
2					
3	*Regression statistics*				
4	Multiple R	0.778555846			
5	R Square	0.606149205			
6	Adjusted R Square	0.578017005			
7	Standard Error	0.739852874			
8	Observations	16			
9					
10		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
11	Intercept	3.505122916	0.303616403	11.54458	1.53E-08
12	nea	-0.003441487	0.00074141	-4.64182	0.000381
13					

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.779[a]	.606	.578	.7399

[a] Predictors: (Constant), nea

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.505	.304		11.545	.000
	nea	−.003	.001	−.779	−4.642	.000

[a] Dependent Variable: fat

Figure 4.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 4.2 Least-squares regression for the nonexercise activity data: output from a graphing calculator, three statistical programs, and a spreadsheet program.

APPLY YOUR KNOWLEDGE

4.1 Don’t drink and drive. The Department of Motor Vehicles warns of the effect of drinking alcohol on blood alcohol content (BAC, in percent of volume). One “drink” is defined as 12 ounces (oz) of beer, or 4 oz of table wine, or 1.25 oz of 80-proof liquor.[1]

a. The impact of a single drink on a person’s BAC depends on sex and weight. For a 160-pound man, the blood alcohol content as a function of the number of drinks consumed in one hour can be expressed by the regression line

$$\text{BAC} = 0.023 \times \text{number of drinks}$$

$$\text{BAC} = 0.023 \times \text{number of drinks}$$

Say in words what the slope of this line tells you. Why do you think the intercept is zero?

b. For a woman weighing 120 pounds, BAC increases by 0.038, on average, for each drink taken in one hour. What is the equation of the regression line for predicting BAC from number of drinks for a 120-pound woman?

c. Find the predicted BAC for a 160-pound male who consumed 4 drinks over the last hour and for a 120-pound woman having 2 drinks in one hour. Explain why these values may not be the individuals' actual blood alcohol contents.

4.2 Chickadee alarm calls. The black-capped chickadee (*Poecile atricapilla*) is a small songbird that uses complex language to communicate about food sources and predator threats. Researchers recorded chickadee vocalizations in an aviary when the birds were presented with predators of various sizes. Exercise 3.5 (page 74) gives the average number of D notes per chickadee warning call for each type of predator, along with the predator wingspan (in centimeters) for that experiment. As you found out in Exercise 3.5, the relationship between these two variables is linear.

a. Find the least-squares regression line for predicting y from x. Make a scatterplot with the regression line shown on the plot.

b. Explain in context what the slope of the regression line tells us.

4.3 Coral reefs. Exercise 3.2 discusses a study in which scientists examined data on mean summer sea-surface temperatures (in degrees Celsius) and mean coral growth (in millimeters per year) over a several-year period at locations in the Red Sea. Here are the data:

Temperature ($^\circ$C)	29.68	29.87	30.01	30.25	30.47	30.65	30.90
Growth (mm/y)	2.63	2.57	2.67	2.60	2.47	2.39	2.25

a. Use your calculator to find the mean and standard deviation of both sea-surface temperature x and growth y as well as the correlation r between x and y. Use these basic measures to find the equation of the least-squares line for predicting y from x.

b. Enter the data into your software or calculator and use the regression function to find the least-squares line. The result should agree with your work in part a up to roundoff error.

Amar and Isabelle Guillen-Guillen Photography/Alamy Images

FACTS ABOUT LEAST-SQUARES REGRESSION

One reason for the popularity of least-squares regression lines is that they have many convenient special properties. Here are some facts about least-squares regression lines. Facts 2, 3, and 4 are special properties of least-squares regression—they are not true for other methods of fitting a line to data.

Fact 1. The distinction between explanatory and response variables is essential in regression. We want to express y as a function of x. Least-squares regression makes the distances of the data points from the line small only in the y direction. If we reverse the roles of the two variables, we get a different least-squares regression line.

Fact 2. There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b = r\frac{s_y}{s_x}$$

You see that **the slope and the correlation always have the same sign.** For example, if a scatterplot shows a positive association, then both b and r are positive. They should not be confused, however. The slope measures the average change in y for a one-unit change in x, whereas the correlation measures the strength and direction of the linear relationship.

Fact 3. The least-squares regression line always passes through the point $(\bar{x}, \bar{y})$ on the graph of y against x. Remember that the intercept is $a = \bar{y} - b\bar{x}$. So, when $x = \bar{x}$, $y = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}$.

Fact 4. The correlation r describes the strength of a straight-line relationship. In the regression setting, **the square of the correlation,** r^2**, is the fraction of the variation in the values of** y **that is explained by the least-squares regression of** y **on** x**.**

INTERPRETING r^2

The squared value of the correlation coefficient, r2r^2, tells us what fraction of the variation in the yy variable is explained by the least-squares regression model.

When there is a linear relationship, some of the variation in yy is accounted for by the fact that as xx changes, the predicted response y^$\hat{y}$ is pulled along with it. This is the variation in yy that is explained by the regression model. Although we will skip the algebra,[2] it’s possible to write

r2=variation in y^ as x pulls it along the linetotal variation in observed values of y

$$r^2 = \frac{\text{variation in } \hat{y} \text{ as } x \text{ pulls it along the line}}{\text{total variation in observed values of } y}$$

or, equivalently,

r2=total variation in observed values of y – variation of the residualstotal variation in observed values of y

$$r^2 = \frac{\text{total variation in observed values of } y - \text{variation of the residuals}}{\text{total variation in observed values of } y}$$

Unless all the data points lie exactly on a line, there is additional variation in the actual response yy that appears as the scatter of points above and below the line. Recall that the least-squares method gives us, by definition, the line with the smallest possible sum of squared vertical distances from the points to the line. These squared vertical distances are the still-unexplained variation in yy.

EXAMPLE 4.3 Using r2r^2

For the NEA data, r=-0.7786$r = -0.7786$ and r2=0.6062$r^2 = 0.6062$. Nearly 61% of the variation in fat gain is accounted for by the linear relationship with change in NEA. The other 39% is individual variation among subjects that is not explained by the linear relationship.

By comparison, Figure 3.1 (page 71) shows a stronger linear relationship in which the points are more tightly concentrated along a line. Here, r=0.9448$r = 0.9448$ and r2=0.8927$r^2 = 0.8927$. Approximately 89% of the variation in manatee deaths from collisions with powerboats in a given year is explained by the number of powerboats registered that

> year. Only about 11% is variation among years with the same number of powerboat registrations.

When you report a regression, give r^2 as a measure of how successful the regression was in explaining the response. All the outputs in Figure 4.2 include r^2, either in decimal form or as a percent (look for “R Square,” “R-Sq,” or “Multiple R-squared”). Perfect correlation ($r = -1$ or $r = 1$) means that the points lie exactly on a line. In that case, $r^2 = 1$, and all the variation in one variable is accounted for by the linear relationship with the other variable. If $r = -0.7$ or $r = 0.7$, then $r^2 = 0.49$, and about half the variation is accounted for by the linear relationship.

Many students are puzzled by the use of both r *and* r^2 *in linear regression.* Almost anything can be computed, such as $1/(r + 2)$ or $r^3 + 10$, but few of these values will be meaningful in a practical sense. The values of r and r^2 are commonly cited because both are informative and easy to interpret: r is a measure of the strength and direction of a linear relationship, whereas r^2 represents the fraction of the variation in y that can be explained by the linear regression model. They are related yet different pieces of information.

CAUTION

APPLY YOUR KNOWLEDGE

4.4 Tara gum. Food additives, such as the plant-based extract tara gum, are used in industrial food preparation as thickening agents to stabilize emulsions and create an impression of thickness and creaminess without adding extra calories. The effect of tara gum on the viscosity (in millipascal-seconds, mPa.s) of fruit yogurt drinks was assessed for various concentrations of tara gum (in grams per 100 grams of drink, g/100 g). Here are the findings:[3]

Concentration (g/100 g)	0.00	0.03	0.06	0.09	0.13	0.16	0.19	0.22
Viscosity (mPa.s)	11.7	20.9	25.7	31.4	42.4	53.8	60.8	81.5
Concentration	0.25	0.28	0.31	0.34	0.38	0.41	0.44	0.47

(g/100 g)								
Viscosity (mPa.s)	105.4	126.5	150.3	181.2	210.8	216.0	303.5	316.2

a. Make a scatterplot of the data. Find the least-squares regression line for predicting drink viscosity from tara gum concentration, and add this line to your plot. Why should we *not* use the regression line for prediction in this setting?
b. Even regression lines that make no practical sense obey Facts 2, 3, and 4. Use the equation of the regression line you found in part a to show that when x$\bar{x}$ is the mean tara gum concentration, the predicted viscosity y^$\hat{y}$ is the mean of the observed viscosities. Always plot your data before using regression.

Monte dei Paschi di Siena A.p.A.

4.5 Magellanic penguins. Example 3.6 (page 73) described a study of Magellanic penguins’ breeding. Figure 3.3 (page 73) shows the relationship between starvation and fledging in the Punta Tombo colony for the 28 years between 1983 and 2010. There are two outliers on this scatterplot, which the researchers excluded from their analysis because of those years’ highly unusual climatic conditions. Here is a quote from the publication: “Starvation strongly affected reproductive success. When a higher percentage of chicks starved, a lower percentage fledged (R2=0.70$R^2 = 0.70$ excluding 1991 and 1999).”

a. Interpret the quoted value 0.70 in the context of this study.
b. What is the value of the correlation between percent of chicks starved and percent of chicks fledged?

4.6 Understanding r2r^2. Exercise 3.8 (page 81) described a study of emotional distress as a function of time since a breakup. The scatterplot you made showed a clear linear pattern. As a pedagogical exercise, we will find the value of r2r^2 ina step-by-step manner to understand what it represents. Here are the data from the study and some corresponding calculations:

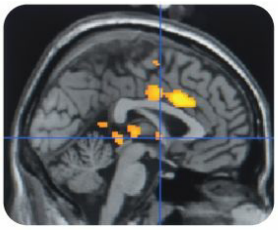
xx	**y**y	**y^**$\hat{y}$	**y −y^**$y-\hat{y}$	**y −y¯**$y-\bar{y}$
0	3.3	3.2763	0.0237	0.325
2	3.1	3.1424	−0.0424	0.125
6	2.9	2.8746	0.0254	−0.075
10	2.6	2.6068	−0.0068	−0.375

a. The mean of the four average distress values is y¯=2.975.$\bar{y} = 2.975.$ Compute the square of the four differences (y-y¯$y-\bar{y}$) and add them up. The value you get, ∑(y-y¯)2,$\sum(y-\bar{y})^2$, corresponds to the total variation in yy. (Notice that this is the numerator of the equation for the variance of yy from Chapter 2.)
b. Now square the regression residuals (y-y^$y-\hat{y}$) and add them up. The value you get, ∑(y-y^)2,$\sum(y-\hat{y})^2$, is the variation in yy that is still unexplained after fitting the linear model. Notice that this value is smaller than the value you computed in part a.
c. What fraction of the variation in yy is explained by the linear model? Subtract the value you obtained in part b from the value you obtained in part a, then divide this difference by the value you obtained in part a. As with an item on sale (10% off or 50% off), this calculation tells you how much lower the variability of the regression residuals is relative to the total variability of yy. This value is r2r^2.
d. Use technology to obtain the value of r2r^2. The result should agree with your own computation, up to roundoff error. Interpret this value in the context of the problem.

OUTLIERS AND INFLUENTIAL OBSERVATIONS

One of the first principles of data analysis is to look for an overall pattern and also for striking deviations from the pattern. A regression line describes the overall pattern of a linear relationship between an explanatory variable and a response variable. We can identify deviations from this pattern by looking at the scatter of the data points about the regression line. When examining scatterplots, look for points with large residuals—outliers of the relationships—and any other unusual observations.

EXAMPLE 4.4 I feel your pain



Philippe Psaila/Science Source

"Empathy" means being able to understand what others feel. To see how the brain expresses empathy, researchers recruited 16 couples in their midtwenties who were married or had been dating for at least two years. They zapped the man's hand with an electrode while the woman watched and while they measured the activity in several parts of the woman's brain that would respond to her own pain. Brain activity was recorded as a fraction of the activity observed when the woman herself was zapped with the electrode. The women also completed a psychological test that measures empathy. The research objective is to see if women who score higher in empathy respond more strongly when their partner has a painful experience. Here are data for one brain region:[4]

Subject	1	2	3	4	5	6	7	8
Empathy score	38	53	41	55	56	61	62	48
Brain activity	−0.120	0.392	0.005	0.369	0.016	0.415	0.107	0.506
Subject	9	10	11	12	13	14	15	16

Empathy score	43	47	56	65	19	61	32	105
Brain activity	0.153	0.745	0.255	0.574	0.210	0.722	0.358	0.779

Figure 4.3 is a scatterplot of these data, with empathy score as the explanatory variable xx and brain activity as the response variable yy. The plot shows a positive association. That is, women who score higher in empathy do indeed react more strongly to their partners' pain. The overall pattern is moderately linear, with correlation r=0.515$r = 0.515$.

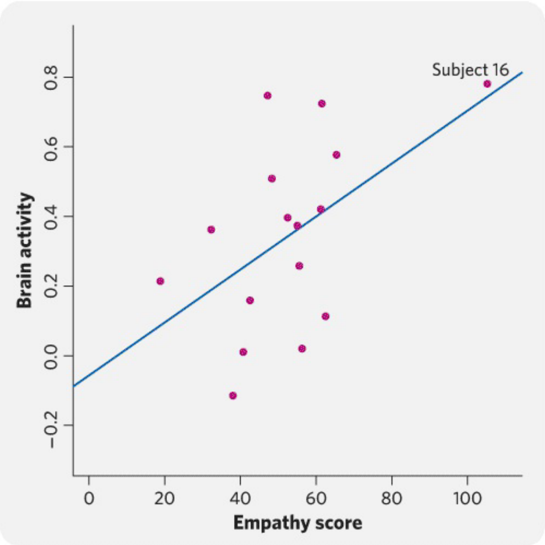


Figure 4.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 4.3 Scatterplot of activity in a region of the brain that responds to pain versus score on a test of empathy, for Example 4.45. Brain activity is measured as the subject watches her partner experience pain. The line is the least-squares regression line. Subject 16 is an outlier in the x direction.

The solid line on the plot is the least-squares regression line of brain activity on empathy score. Its equation is

y^=-0.0578+0.00761x

$$\hat{y} = -0.0578 + 0.00761x$$

Figure 4.3 shows one unusual observation. Subject 16 is an outlier in the xx direction, with an empathy score 40 points higher than the score for any other subject. However, this point fits the overall pattern and, therefore, is not an outlier of the relationship. Does this unusual data point have any influence on the correlation and regression equation?

INFLUENTIAL OBSERVATIONS

An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation.

The result of a statistical calculation may be of little practical use if it depends strongly on one or more influential observations.

Points that are outliers in either the x or y direction of a scatterplot are often influential for the correlation. Points that are outliers in the x direction are often influential for the least-squares regression line.

EXAMPLE 4.5 An influential observation?

Because of this point's extreme position on the empathy scale x, Subject 16 has a strong influence on the correlation. Dropping this point reduces the correlation from $r = 0.515$ to $r = 0.331$. We say that Subject 16 is *influential* for calculating the correlation.

Is this observation also influential for the least-squares line? Figure 4.4(a) shows that it is not. The regression line calculated without Subject 16 (dashed) differs little from the line that uses all the observations (solid). The reason that this outlier in the x direction has little influence on the regression line is that it is not an outlier of the relationship and lies close to the dashed regression line calculated from the other observations.

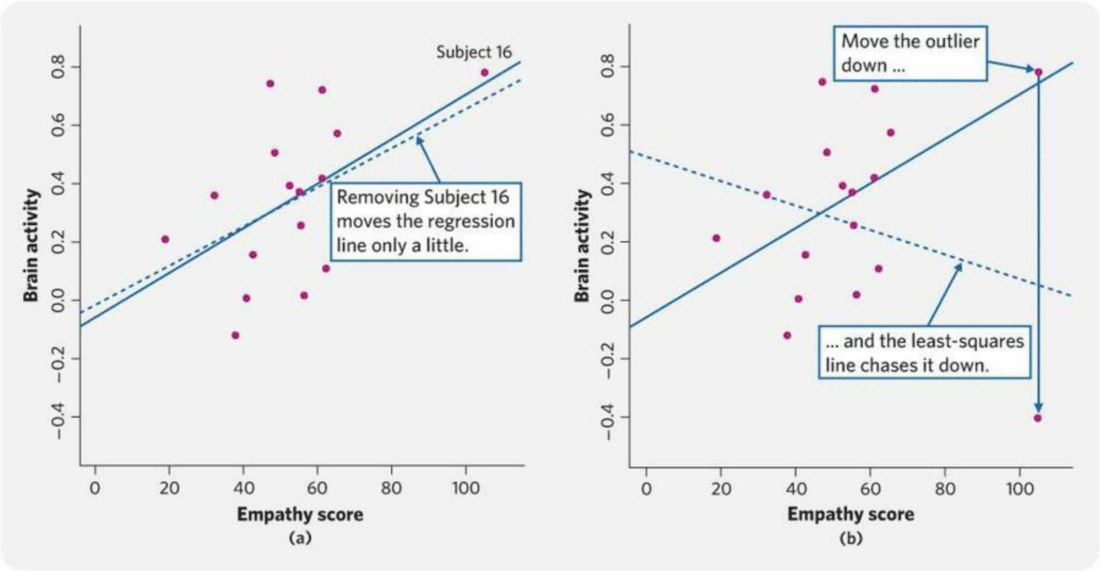


Figure 4.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 4.4 Variations on Figure 4.3.(a) for Example 4.5. (a) Subject 16 is an outlier in the x direction. It is not influential for least-squares regression, because removing it moves the regression line (dashed line) only a little. (b) An outlier in the x direction pulls the least-squares line to itself when there are no other observations with similar values of x to hold the line in place. Altering Subject 16’s data to move its value down pulls the regression line down (dashed line).

APPLET

Now let’s try an experiment. Suppose we pull Subject 16’s point in the scatterplot straight down. This altered point is now an outlier of the relationship in addition to being an outlier in the x direction. What happens to the regression line? Figure 4.4(b) shows the result. The dashed line is the regression line with Subject 16’s altered data. Because there are no other points with similar x-values, the dashed line chases the outlier. An outlier in x pulls the least-squares line toward itself. If the outlier in the x direction does not lie close to the line calculated from the other observations, it will be influential. You can use the *Correlation and Regression* applet to animate any scatterplot and perform your own experiment.

CAUTION

We did not need the distinction between outliers and influential observations in Chapter 2. When studying one variable by itself, an outlier that lies far above the other values pulls up the mean $\bar{x}$ toward it and, therefore, is always influential. *In the regression* *setting, however, not all outliers are influential.*

APPLY YOUR KNOWLEDGE

4.7 Magellanic penguins, continued. Here are the data for the study of Magellanic penguins in Example 3.6 and Exercise 4.5. The values in italics correspond to the two years with highly unusual climatic conditions that the researchers excluded from their analysis.

Percent starved	Percent fledged	Percent starved	Percent fledged	Percent starved	Percent fledged	Percent starved	Percen fledged
39.5	40.8	54.0	20.3	27.4	27.1	47.2	34.7
20.9	43.7	66.0	24.2	38.3	27.8	29.0	37.1
25.9	45.5	28.1	54.2	32.4	57.3	18.6	66.4
44.0	45.8	55.3	26.0	39.0	44.4	39.2	42.6
66.0	1.7	16.8	55.7	28.6	58.2	46.0	32.8
86.4	6.8	46.9	27.2	11.8	60.5	*20.9*	*11.2*
69.8	8.6	40.8	26.9	28.6	31.4	*34.0*	*10.5*

a. Are the two outliers (values in italics) influential for the correlation? Obtain r with and without these two data points.

b. Are the two outliers influential for regression? Obtain the equation of the least-squares regression line with and without these two data points. Add both regression lines to the same scatterplot, as in Figure 4.4(a).

c. What are the two excluded data points? Are they outliers in the x direction? In the y direction? Are they outliers of the relationship? Explain how this is relevant to your conclusions in parts a and b.

4.8 Chickadee alarm calls, continued. In Exercise 4.2 you obtained the regression line for the relationship between the average number of D notes per chickadee warning call and a predator’s wingspan (in centimeters). The researchers also recorded chickadee calls in response to the introduction of a bobwhite, a harmless quail. The bobwhite has a wingspan of 48.1 cm, and the chickadees responded to its presence with an average of 1.74 D notes per call.

a. Make a scatterplot of the data with this additional point. Find the equation for the new regression line when the bobwhite data point is included and draw it on your graph.

b. Is the bobwhite data point an outlier of the relationship? Is it an outlier in the distribution of the x variable? Is it an outlier in the distribution of the y variable? Is it influential? Explain why the bobwhite data point is interesting as a comparison but should not be included to describe the relationship between D notes per warning call and predator wingspan.

Frank Leung/iStockphoto

WORKING WITH LOGARITHM TRANSFORMATIONS[i]

Logarithm transformations are quite frequent in biology to study problems ranging from population growth to animal physiology. A regression line can be computed using the transformed data, and prediction within the range of the data can be performed as with simple linear relationships. The difference is that we will need to transform the predicted y back into its original (nonlogarithmic) unit.

EXAMPLE 4.6 Using a regression line: transformed data

Humans have large brains compared with other mammals, but they are also a relatively large mammal species. How does an animal’s brain relate to its body size? Figure 4.5(a) shows the scatterplot of brain weight in grams against body weight in kilograms for 96 species of mammals.[5] Figure 4.5(b) shows the scatterplot for the same data after a logarithm transformation. The line is the least-squares regression line for predicting y from x.

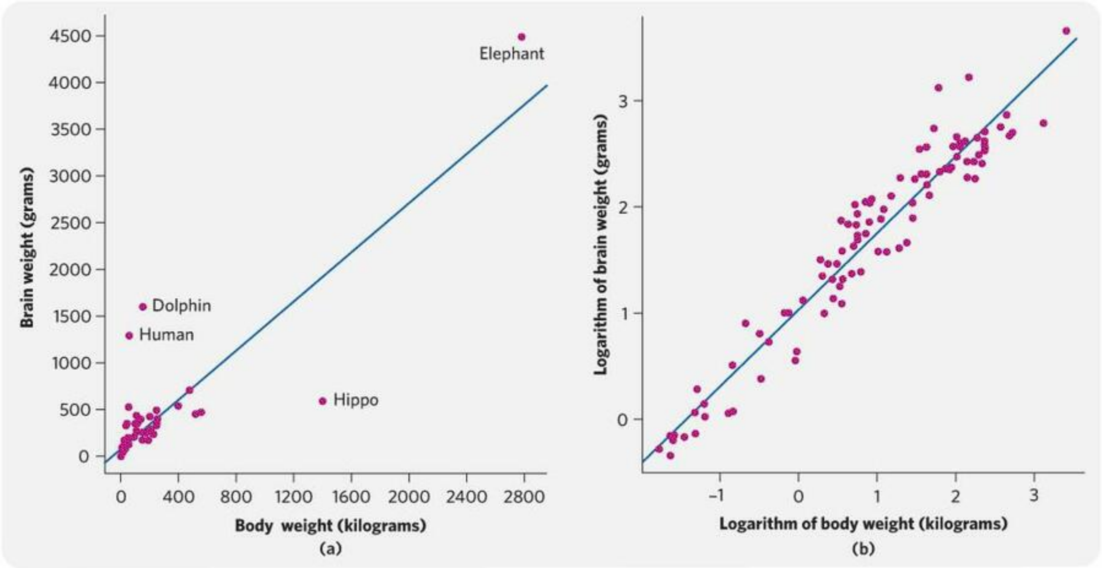


Figure 4.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 4.5 (a) Scatterplot of body weight and brain weight for 96 species of mammals in the original numerical values of kilograms and grams, for Example 4.6. (b) Scatterplot of the data in (a) using the logarithm of brain weight and the logarithm of body weight.

The regression line for the untransformed data in Figure 4.5(a) is clearly unsatisfactory because most mammals are so small compared to elephants. The correlation between brain weight and body weight is r=0.86,$r = 0.86$, but this value is misleading. If we remove the elephant data point, the correlation for the other 95 species drops to r=0.50.$r = 0.50$. Humans, dolphins, and hippos are also clear outliers of the relationship.

The scatterplot using transformed data in Figure 4.5(b) has no obvious outliers and is clearly linear, with correlation r=0.96.$r = 0.96$. The least-squares regression for predicting the logarithm of brain weight from the logarithm of body weight has the equation

log(brain weight)=1.01+[0.72 × log(body weight)]

$$\log(\text{brain weight}) = 1.01 + [0.72 \times \log(\text{body weight})]$$

For a mammal species weighing, on average, 100 kg, the logarithm is log(100) = 2, and the predicted logarithm of brain weight is

log(brain weight)=1.01+(0.72 × 2)=2.45

$$\log(\text{brain weight}) = 1.01 + (0.72 \times 2) = 2.45$$

To undo the logarithm transformation, recall that for common logarithms with base 10, y=10log(y)$y = 10^{\log(y)}$. Thus, the predicted brain weight, in grams, for the 100-kg species is

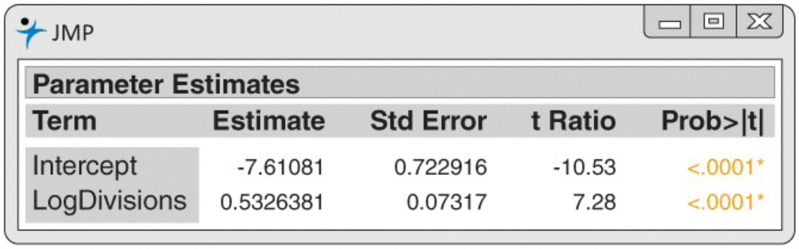
brain weight=102.45=282 g

$$\text{brain weight} = 10^{2.45} = 282\dagger\ \text{g}$$

Examine Figure 4.5 and verify that this prediction makes sense on both scatterplots, although the scatterplot of transformed data is much easier to use.

APPLY YOUR KNOWLEDGE

4.9 Predicting cancer. Exercise 3.33 (page 88) examined the relationship between the total number of stem cell divisions in the lifetime of a given tissue and the lifetime risk of cancer for that tissue, based on 31 types of cancers for which this information is known in the U.S. population. After the data from Table 3.6 (page 89) are converted to a logarithm scale (base 10), a scatterplot of the results shows a linear pattern. Figure 4.6 gives part of the regression output for the log-transformed data using statistical software.



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-7.61081	0.722916	-10.53	<.0001*
LogDivisions	0.5326381	0.07317	7.28	<.0001*

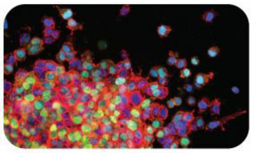
Figure 4.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 4.6 JMP regression output for the cancer study of Exercise 4.9.

a. Rounding the coefficients to four decimal places, give the equation of the least-leastsquares regression line modeling LogRisk (yy) as a function of LogDivisions (xx).

b. Using this equation, predict the log of lifetime cancer risk y^$\hat{y}$ for a tissue with a lifetime total of 1 billion (109$10^9$) stem cell divisions. Be sure to first find the value of xx that corresponds to 1 billion divisions.

c. Following the steps described in Example 4.6, undo the logarithm transformation to find the predicted lifetime cancer risk for a tissue with a lifetime total of 1 billion stem cell divisions.



National Cancer Institute/Winship
Cancer Institute of Emory
University

CAUTIONS ABOUT CORRELATION AND REGRESSION

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, however, you must be aware of their limitations. You already know:

- *Correlation and regression lines describe only linear relationships.* You can do the calculations for any relationship between two quantitative variables, but the results are useful only if the scatterplot shows a linear pattern.
- *Correlation and least-squares regression lines are not resistant.* Always plot your data and look for potentially influential observations.

Here are two more things to keep in mind when you use correlation and regression.

Beware extrapolation. Suppose that you have data on a child's growth between 3 and 8 years of age. You find a strong linear relationship between age x and height y. If you fit a regression line to these data and use it to predict height at age 25 years, you will predict that the child will be more than 8 feet tall. Growth slows down and then stops at maturity, so extending the straight line to adult ages is foolish. *Few relationships are linear for all values of* x. *Don't make predictions for values of* x *outside the range that actually appears in your data.*

EXTRAPOLATION

Extrapolation is the use of a regression line for prediction well outside the range of values of the explanatory variable x that you used to obtain the line. Such predictions are often not accurate.

Beware the lurking variable. Another caution is even more important: *The relationship between two variables can often be understood only by* *taking other variables into account. Lurking variables* can make a correlation or regression misleading.

LURKING VARIABLE

A **lurking variable** is a variable that is not one of the explanatory or response variables in a study, yet may influence the interpretation of relationships among those variables.

You should always think about possible lurking variables before you draw conclusions based on correlation or regression. You should also consider that an association can sometimes be markedly different for different groups. The failure to recognize such issues can lead to an incorrect interpretation of the association between two variables.

SPURIOUS CORRELATION

A series of research papers described a puzzling, unexplainable positive correlation between the prevalence of doctors and infant mortality rate in developed countries. Notice, however, that both doctor prevalence and infant mortality rate are ratios of counts divided by the population size. In fact, a later paper reported a significant negative association between the number of physicians and infant mortality, after adjusting for population size. The first series of papers had failed to take into account population size as a lurking variable.

EXAMPLE 4.7 Alcohol consumption and longevity

A number of studies have found that individuals who consume moderate amounts of alcohol are more likely to live longer. News reports of the potential health benefit of moderate alcohol consumption often follow the publication of such studies. But can we legitimately conclude that alcohol consumption itself is a cause of greater longevity?

Researchers reviewed 87 different scientific studies examining the association between alcohol consumption and longevity.[6] They found that moderate drinkers tended to be healthier not necessarily because of their alcohol consumption, but because they tended to be better educated and more affluent. A person’s socioeconomic situation produces a number of lurking variables that explain, at least in part, why longevity is associated with alcohol consumption.

In addition, people who do not consume alcohol may have radically different reasons to do so. Some people may choose to abstain from alcohol as a lifelong lifestyle decision, whereas others may have been compelled to stop drinking because of poor health. These individuals are likely to lower the average longevity of all study participants who do not drink alcohol, so that the group with moderate drinking appears to have the better outcome.

Teasing out the relative influence of multiple factors in complex environments can be very challenging. Identifying potential confounding variables requires some expertise on the topic studied or at the very least a healthy dose of common sense. No statistical software will do this automatically for you. Planning a **multivariate** study and analysis that includes potential lurking variables in addition to your main variables is an important part of statistics necessary to reach sound conclusions.

EXAMPLE 4.8 Men’s and women’s physiology

High blood pressure and arterial stiffness are risk factors for cardiovascular disease, the leading cause of death worldwide. Understanding the physiology of cardiovascular health is important for developing preventive and curative approaches to the disease.

The sympathetic nervous system is known to affect vasoconstriction. A research team investigated the relationship between sympathetic nerve activity (measured in number of bursts per 100 heartbeats) and an indicator of arterial stiffness (augmented aortic pressure, in millimeters of mercury, mm Hg) in 44 healthy young adults.[7] The results, displayed in Figure 4.7(a), show no clear relationship between sympathetic nerve activity and the arterial stiffness indicator (r=-0.17$r = -0.17$).

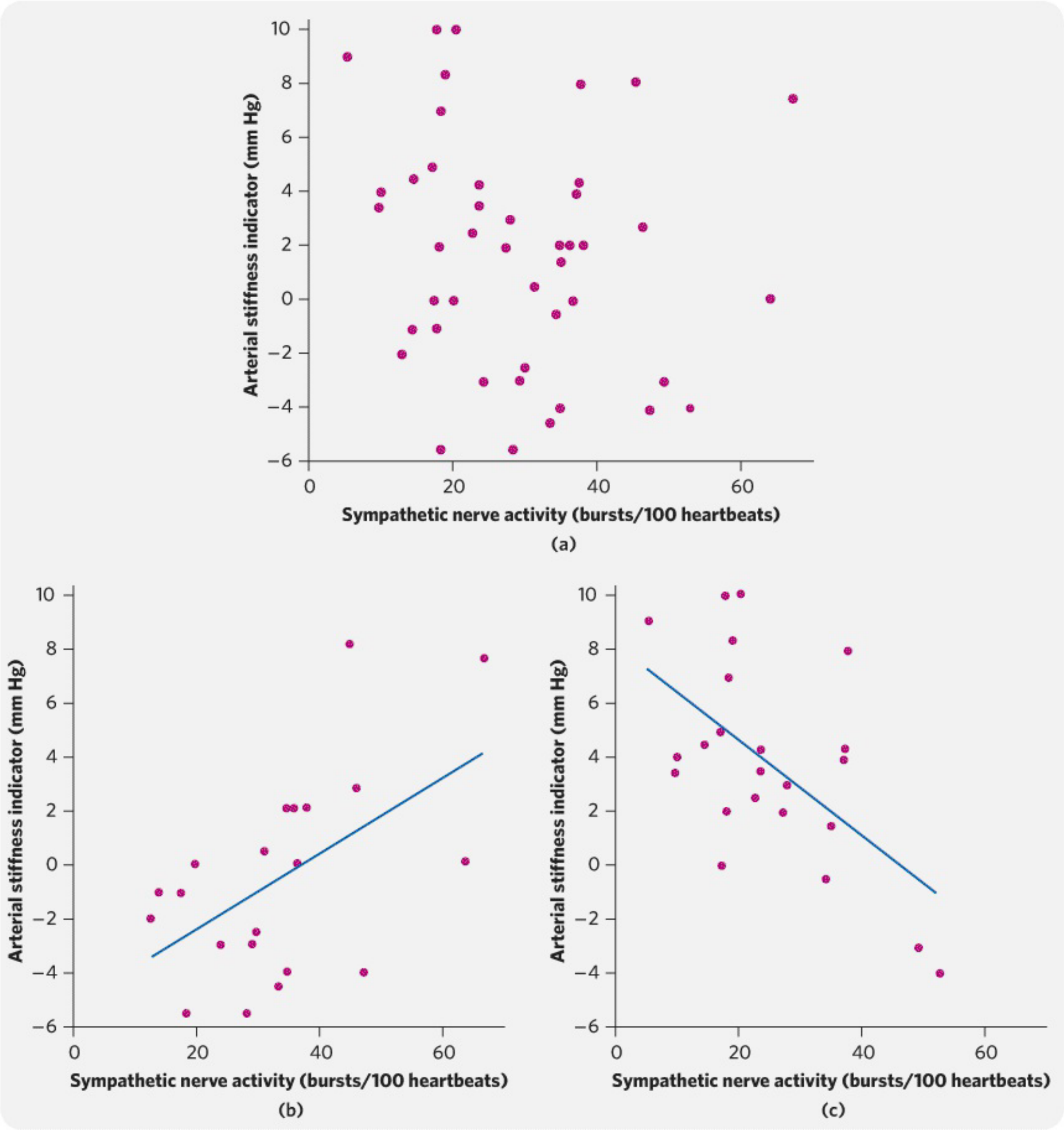


Figure 4.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 4.7 (a) Scatterplot of sympathetic nerve activity and an indicator of arterial stiffness in 44 healthy young adults, for Example 4.8. Notice the lack of any obvious association. (b)–(c) Scatterplot of the data in (a) after separating the data for men (b) and for women (c). Notice the opposite trends for men (positive linear association) and for women (negative linear association).

Men and women have a somewhat different cardiovascular physiology, so the researchers decided to examine their findings separately for men and for women. The results are displayed in Figure 4.7(b) and (c), respectively. To the researchers' surprise, a moderate linear relationship was noticeable in both cases, but the relationship was

> positive for men (r=0.53$r = 0.53$) and negative for women (r=-0.58 $r = -0.58$).
>
> In this case, combining data from men and women created the *appearance* of no relationship between sympathetic nerve activity and the arterial stiffness indicator. Gender is a lurking variable that drastically affects the relationship between the two quantitative variables studied.

Lastly, avoid the temptation to study the relationship between any two variables just to see if something might come up. For instance, the worldwide AIDS epidemic and Internet use have both exploded since the 1980s, so you could expect to find a positive association between both variables. Of course, this association is purely coincidental and meaningless. Example 4.9 offers a humorous illustration of this point. Always remember that statistics is the analysis of data in context. This warning is especially important in the age of “big data,” when our capacity to collect enormous amounts of information sometimes outpaces our ability to make sense of it.

APPLY YOUR KNOWLEDGE

4.10 The endangered manatee. Table 3.1 (page 70) gives data on powerboats registered in Florida and manatees killed by collisions with powerboats from 1977 to 2016. Figure 3.1 (page 71) shows a strong positive linear relationship between these two variables.

a. Find the equation of the least-squares line for predicting manatees killed by powerboats from thousands of boats registered. Because the linear pattern is so strong, we expect predictions from this line to be quite accurate—but only if conditions in Florida remain similar to those of the past 40 years.
b. Manatees are an endangered species. Imagine that Florida limited the powerboat registrations in a given year to 500,000. Predict the number of manatees killed by powerboats in such a year. (Pay attention to the fact that registrations in this data set

are recorded in thousands.) Explain why we can trust this prediction.

c. Now imagine that Florida took a drastic measure and limited the number of powerboat registrations to 200,000 per year. Explain why you shouldn't use the regression equation you found in part a to predict the number of manatee deaths by powerboat in this condition. Make the calculation anyway and comment on how this prediction indeed makes no sense.

d. Now imagine that powerboats became even more popular in Florida and that there were 1.5 million powerboats registered in a year. Make the prediction regardless of its validity. Explain why you shouldn't use the regression equation you found in part a to make this prediction even though it doesn't lead to a completely senseless result.

4.11 New dosage for sleep drugs. In 2013, the U.S. Food and Drug Administration made an announcement regarding widely used sleep drugs containing the active ingredient zolpidem. It contained the following statement:[8] "Since women eliminate zolpidem from their bodies more slowly than men, the FDA has notified the manufacturers that the recommended dose should be lowered for women."

a. Drug safety initially had been established by studying zolpidem blood levels as a function of time since ingestion of the drug. State the explanatory and response variables here.

b. A lurking variable was later revealed, leading the FDA to issue a notification. What is this lurking variable? What is the reason cited explaining the impact of this lurking variable?

ASSOCIATION DOES NOT IMPLY CAUSATION

Thinking about lurking variables leads to the most important caution about correlation and regression. When we study the relationship between two variables, we are often interested in whether changes in the explanatory variable *cause* changes in the response variable. *A strong association between two variables is not enough to draw conclusions about cause and effect.* Sometimes an observed association really does reflect cause and effect. A household that heats with natural gas uses more gas in colder months because cold weather requires burning more gas to stay warm. In other cases, an association is explained by lurking variables, and the conclusion that xx causes yy is either wrong or not proved. *The important caution that association does not necessarily imply causation applies whether the variables considered are quantitative or categorical.*

EXAMPLE 4.9 Chocolate and Nobel Prizes

Even reputable publications like the *New England Journal of Medicine* can display a certain sense of humor . . . Right around Nobel Prize season, the journal published a satirical editorial describing an analysis of the correlation between chocolate consumption xx (in kilograms per year per capita) and the number of Nobel laureates yy (per 10 million population) among nations that have ever received a Nobel Prize. A scatterplot in the editorial shows a positive linear trend with r=0.791 $r = 0.791$.[9]

The basic meaning of causation is that by changing xx we can bring about a change in yy. Could a country consume more chocolate to boost its number of Nobel laureates? No. As the editorial points out, chocolate consumption is computed at the population level and there is no indication of how much chocolate Nobel laureates actually consume. This makes it difficult to justify a plausible cause-and-effect mechanism between per-capita chocolate consumption and per-capita number of Nobel laureates.

Correlations such as that in Example 4.9 are sometimes called “nonsense correlations” or “spurious correlations.” In such a case, the correlation is real; what is nonsense is the conclusion that changing one of the variables causes changes in the other. A lurking variable—such as possibly a country’s wealth in Example 4.9—that influences both xx and y y can create a high correlation even though there is no direct connection between xx and yy.

ASSOCIATION DOES NOT IMPLY CAUSATION

An association between an explanatory variable xx and a response variable yy, even if it is very strong, is not by itself good evidence that changes in xx actually cause changes in yy.

EXAMPLE 4.10 Obesity in mothers and daughters

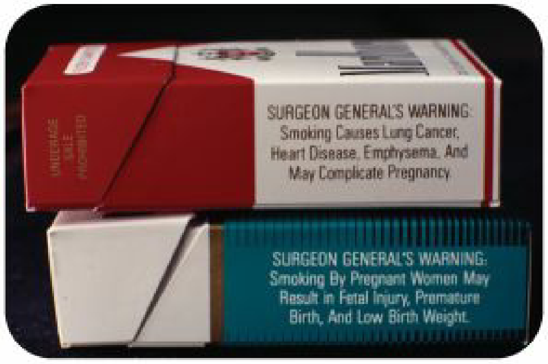
Obese parents tend to have obese children. But is obesity determined genetically? The results of a study of Mexican American girls aged 9 to 12 years are typical. The investigators measured body mass index (BMI), a measure of weight relative to height, for both the girls and their mothers. People with high BMI are labeled obese. The correlation between the BMI of daughters and the BMI of their mothers was r=0.506 $r = 0.506$.[10]

Body type is in part determined by heredity. Daughters inherit half their genes from their mothers. As a consequence, there is the potential for a direct causal link between the BMI of mothers and daughters. However, mothers who are obese may also set an example of little exercise and poor eating habits. It is also possible that both mothers and daughters are affected by the same environmental conditions, such as living in an area with limited access to healthy food (a “food desert”) and no parks or public structures to facilitate exercise. Heredity, behavior, and environmental factors all likely contribute to some extent to the mother–daughter correlation.

The lesson of Example 4.10 is more subtle than just “association does not imply causation.” *Even when direct causation is present, it may not be the whole explanation for a correlation.* You must still worry about lurking variables. Careful statistical studies try to anticipate lurking variables and measure them. The mother–daughter study did measure TV viewing, exercise, and diet. Elaborate statistical analysis can account for the effects of these variables to come closer to the direct effect of mother’s BMI on daughter’s BMI. Even so, this remains a second-best approach to determining causation. When possible, the best way to get good evidence that x causes y is to do an **experiment** in which we change x and keep lurking variables under control. We will discuss experiments in Chapter 7.

When experiments cannot be done, explaining an observed association can be difficult and controversial. Many of the sharpest disputes in which statistics plays a role involve questions of causation that cannot be settled by experiment. Does using cell phones cause brain tumors? Are our extreme levels of carbon dioxide emissions causing global warming? Is human activity triggering a new, planet-wide mass extinction? All these questions have become public issues. All concern associations among variables. And all have this in common: They try to pinpoint cause and effect in a setting involving complex relations among many interacting variables.

EXAMPLE 4.11 Does smoking cause lung cancer?


UNDERAGE SALE PROHIBITED
SURGEON GENERAL'S WARNING: Smoking Causes Lung Cancer, Heart Disease, Emphysema, And May Complicate Pregnancy.
SURGEON GENERAL'S WARNING: Smoking By Pregnant Women May Result in Fetal Injury, Premature Birth, And Low Birth Weight.

James Leynse/Corbis via Getty Images

Despite the difficulties, it is sometimes possible to build a strong case for causation in the absence of experiments. The evidence that smoking causes lung cancer is about as strong as nonexperimental evidence can

> be.
>
> Doctors had long observed that most patients with lung cancer were smokers. Comparison of smokers and “similar” nonsmokers showed a very strong association between smoking and death from lung cancer. Could the association be explained by lurking variables? Might there be, for example, a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated even if smoking had no direct effect on the lungs. How were these objections overcome?

STATS IN YOUR WORLD

CIGARETTE WARNING LABELS

The history of cigarette warning labels illustrates how challenging establishing causation from evidence of association can be. Since first the Surgeon General’s report on smoking and health, there have been four successive warnings on packages sold in the United States: **1965–**“Caution: Cigarette Smoking May Be Hazardous to Your Health.” **1967—**“Warning: Cigarette Smoking Is Dangerous to Health and May Cause Death from Cancer and Other Diseases. “**1969–**“Warning: The Surgeon General Has Determined That Cigarette Smoking Is Dangerous to Your Health.” **1984–**“SURGEON GENERAL’S WARNING: Smoking Causes Lung Cancer, Heart Disease, Emphysema, and May Complicate Pregnancy.” This warning is still in effect, along with three others.

Let’s answer this question in general terms: What are the criteria for establishing causation when we cannot do an experiment?

- *The association is strong*. The association between smoking and lung cancer is very strong.
- *The association is consistent*. Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study

explains the association.

- *Higher doses are associated with stronger responses*. People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.
- *The alleged cause precedes the effect in time*. Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking rates, again with a lag of about 30 years, and has now passed breast cancer as the leading cause of cancer death among women.
- *The alleged cause is plausible*. Experiments with animals show that tars from cigarette smoke do cause cancer.

In 2014, the U.S. Surgeon General released a comprehensive report on the health consequences of smoking on the 50-year anniversary of the first Surgeon General's warning about the health hazards of smoking. This latest report lists 12 cancers (including lung cancer) and numerous chronic diseases causally linked to smoking.[11] The evidence for causation is overwhelming because of the sheer volume of scientific studies supporting the criteria described previously. When attempting to establish causation, however, keep in mind that one observational study cannot provide evidence as strong as the evidence provided by one well-designed experiment. We will examine why in Chapters 6 and 7.

APPLY YOUR KNOWLEDGE

4.12 Zika and microcephaly. Until recently, the Zika virus was a rare and mild public health concern and little was known about it. In early 2015, Brazil reported an epidemic of Zika viral infection. Reports soon followed of unusual numbers of neurological disorders such as microcephaly in newborns, a condition characterized by a very small head and poor brain growth.

a. Explain why we cannot conclude from this information alone that Zika infections cause microcephaly in newborns.

b. In early 2016, the Centers for Disease Control and Prevention (CDC) announced that “Zika causes microcephaly and other birth defects.”[12] The announcement included this sentence: “The report notes that no single piece of evidence provides conclusive proof that Zika virus infection is a cause of microcephaly and other fetal brain defects.” Explain how a conclusion of causality can be reached despite the lack of direct evidence.

4.13 Predicting the risk of breast cancer. A large study of the detailed lifestyle of California Seventh-Day Adventists found that a woman’s level of education is one of the better predictors of her risk of getting breast cancer later in life, with higher levels of education associated with higher risks of breast cancer.[13]

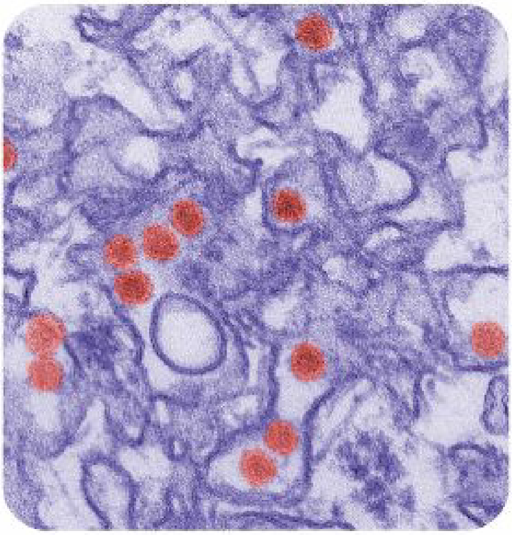
a. Do you think that we should ask women to drop out of school in order to reduce the incidence of breast cancer? Is education a plausible candidate to be a direct cause of breast cancer? Explain.

b. Another variable associated with the risk of breast cancer is the age at which a woman has her first child, with later ages associated with higher cancer rates. Is the age at first child a plausible cause of breast cancer? Explain. How does knowing about the association between age at first child and breast cancer help us understand the association between education and breast cancer?

4.14 Zika and microcephaly, continued. Exercise 4.12 discussed the 2016 CDC announcement that Zika causes microcephaly in newborns. The scientific evidence on which the CDC conclusion is based is available online. Typing DOI:10.1056/NEJMsr1604338 in a search engine will give you a link to the scientific article titled “Zika Virus and Birth Defects: Reviewing the Evidence for Causality.”

a. Table 2 in the article lists a number of criteria that can be used for establishing evidence of causation. Match the criteria from this table to the five criteria we listed for establishing causation in the absence of experiments. Are these five criteria met in the context of Zika infection and microcephaly?

b. Table 2 in the Zika publication lists experimental evidence as one way to establish causation. Is there experimental evidence available in the context of Zika infection and microcephaly? How does this relate to the statement in the CDC announcement that “no single piece of evidence provides conclusive proof that Zika virus infection is a cause of microcephaly?”



CDC/Cynthia Goldsmith

CHAPTER 4 SUMMARY

- A **regression line** is a straight line that describes how a response variable yy changes, on average, as an explanatory variable xx changes. You can use a regression line to **predict** the value of yy for any value of xx by substituting this xx into the equation of the line.
- The **slope** bb of a regression line y^=a+bx$\hat{y}=a+bx$ is the rate at which the predicted response y^$\hat{y}$ changes along the line as the explanatory variable xx changes. Specifically, bb is the change in y^$\hat{y}$ when xx increases by 1 unit.
- The **intercept** aa of a regression line y^=a+bx$\hat{y}=a+bx$ is the predicted response y^$\hat{y}$ when the explanatory variable x=0$x=0$. This prediction is of no statistical interest unless xx can actually take values near 0.
- The most common method of fitting a line to a scatterplot is least squares. The **least-squares regression line** is the straight line y^=a+bx $\hat{y}=a+bx$ that minimizes the sum of the squares of the vertical distances of the observed points from the line.
- The least-squares regression line of yy on xx is the line with slope rsy/sx rs_y/s_x and intercept a=y¯-bx¯ $a=\bar{y}-b\bar{x}$. This line always passes through the point (x¯,y¯)$(\bar{x},\bar{y})$.
- **Correlation and regression** are closely connected. The correlation rr is the slope of the least-squares regression line when we measure both xx and yy in standardized units. The **square of the correlation,** r2,r^2, is the fraction of the variation in one variable that is explained by least-squares regression on the other variable.
- Correlation and regression must be **interpreted with caution. Plot the data** to be sure the relationship is roughly linear and to detect outliers and influential observations.
- Look for **influential observations**, individual points that substantially change the correlation or the regression line. Outliers in the xx direction are often influential for the regression line.

- Avoid **extrapolation,** the use of a regression line for prediction with values of the explanatory variable well outside the range of the data from which the line was calculated.
- **Lurking variables** may explain the relationship between the explanatory and response variables. Correlation and regression can be misleading if you ignore important lurking variables.
- Most of all, be careful not to conclude that a cause-and-effect relationship exists between two variables just because they are strongly associated. **Correlation does not in itself imply causation.** The best evidence that an association is due to causation comes from an **experiment** in which the explanatory variable is directly changed and other influences on the response are controlled.

THIS CHAPTER IN CONTEXT

In this chapter we used the least-squares regression line to describe the straight-line relationship between two quantitative variables (x,yx, y) when such a pattern is seen in a scatterplot (first introduced in Chapter 3). This regression line is a model for an underlying linear pattern and does not imply that yy is an exact function of x.x. Numerous mathematical models can be used to describe univariate, bivariate, or multivariate data. In Chapter 11, we will examine closely one other important mathematical model: Normal distributions.

The equation of the least-squares regression line can be used to make predictions, as long as you are careful not to do so outside the range of values of the explanatory variable that you used to obtain the line. This approach rests on the assumption that the relationship is valid in some broader sense (an issue we will explore more carefully in Chapter 23 when reviewing simple linear regression in the context of statistical inference).

It is important to remember that an observed association does not imply that there is an underlying cause-and-effect relationship between the response and explanatory variables. In fact, a lurking variable may be influencing the interpretation of any association between the response and explanatory variables. This is true regardless of the type of variables studied, quantitative or categorical. In Chapter 5, we examine the association between two categorical variables. In Chapters 6 and 7, we revisit the topic of association versus causation in the context of the limitations arising from data collected in an observational, as opposed to an experimental, setting.

Lastly, in optional Chapter 28, we expand the concept of simple linear regression to include multiple explanatory variables used to describe and predict one response variable (multiple linear regression). We also see how a categorical variable can be coded with zeros and ones to allow its incorporation into the regression model (in multiple linear regression and in logistic regression).

CHECK YOUR SKILLS

4.15 Figure 4.8 is a scatterplot of tree frog mating-call frequency (in notes per second) against outside temperature (in degrees Celsius, °C$^{\circ}\mathrm{C}$) for 16 American bird-voiced tree frogs.[14] The line is the least-squares regression line for predicting call frequency from temperature. If another frog of this species initiates a mating call when the outside temperature is 24°C$24^{\circ}\mathrm{C}$, you predict the call frequency (in notes/s) to be close to

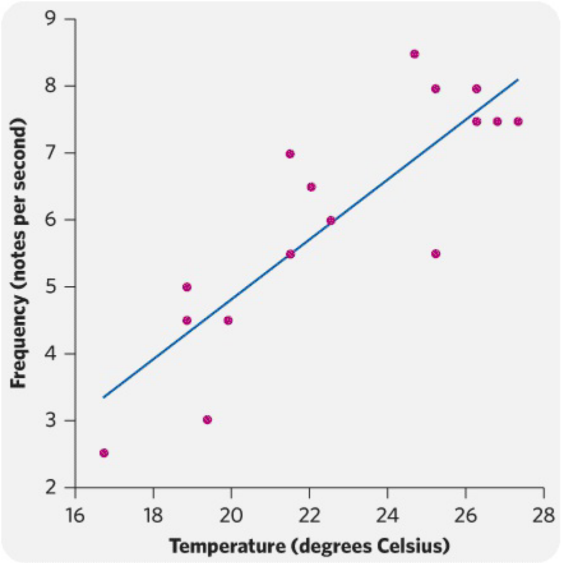


Figure 4.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 4.8 Outside temperature and frequency of mating calls for 16 tree frogs, for Exercise 4.15.

a. 5.
b. 7.
c. 9.

4.16 The slope of the line in Figure 4.8 is closest to (in notes/s/°$\mathrm{notes/s/}^{\circ}$C)

a. –0.5
b. 0.
c. 0.5.

4.17 In the frog mating-call data of Figure 4.8, temperature explains approximately 75% of the variability in call frequency. What is the approximate correlation between call frequency and temperature?

a. 0.5

b. 0.75

c. 0.85

4.18 Given that the relationship between temperature and call frequency in Figure 4.8 is clearly linear and that temperature explains approximately 75% of the variability in call frequency, you conclude that

a. temperature may influence mating-call frequency.

b. temperature causes 75% of the variation in mating-call frequency.

c. temperature and mating-call frequency are unrelated.

4.19 For a class project, you measure the weight in grams (g) and the tail length in millimeters (mm) of a group of mice. The equation of the least-squares line for predicting tail length from weight is

$$\text{predicted tail length}=20+3\times\text{weight}$$

$\text{predicted tail length} = 20 + 3 \times \text{weight}$

How much (on average) does tail length increase for each additional gram of weight?

a. 3 mm

b. 20 mm

c. 23 mm

4.20 According to the regression line in Exercise 4.19, what is the predicted tail length for a mouse weighing 18 g?

a. 74 mm

b. 54 mm

c. 34 mm

4.21 If you had measured the tail length in Exercise 4.19 in centimeters instead of millimeters (1 centimeter = 10 millimeters), what would the slope of the regression line be?

a. 3/10 = 0.3

b. 3
c. $(3)(10) = 30$

4.22 Because elderly people may have difficulty standing straight to have their heights measured, a study looked at predicting overall height from height to the knee. Here are the data (in centimeters, cm) for five elderly men:

Knee height x (cm)	57.7	47.4	43.5	44.8	55.2
Overall height y (cm)	192.1	153.3	146.4	162.7	169.1

Use your calculator or software: What is the equation of the least-squares regression line for predicting height from knee height?

a. $\hat{y} = 2.4 + 44.1x$
b. $\hat{y} = 44.1 + 2.4x$
c. $\hat{y} = -2.5 + 0.32x$

4.23 Marco’s mom measured his height in centimeters (cm) every few months when he was between ages 4.5 and 7.5. The recorded values lie close to the line whose equation is $y = 80.46 + 6.25x$. Using this equation to find out whether Marco will be taller than his mom by age 14 (she is 168 cm tall) is an example of

a. an influential point.
b. prediction within range.
c. extrapolation.

4.24 Across American cities, the number of pharmacists and the number of deaths in a given year are positively associated. What can be concluded from this association?

a. Pharmacists sell dangerous drugs, therefore more pharmacists lead to more deaths.
b. Patients near death are prescribed lots of medication, therefore more deaths means more pharmacists are needed.
c. Larger cities tend to have both more deaths and more pharmacists, therefore the association may not be causal.

CHAPTER 4 EXERCISES

4.25 More on tara gum. The study described in Exercise 4.4 also examined the impact of tara gum on the perceived creaminess of fruit yogurt drinks. Participants were asked to rate creaminess on a scale of 0 to 100 (with 0 meaning “not at all”) for fruit yogurt drinks containing various concentrations of tara gum ranging from 0 to 0.47 g per 100 g of drink. The observed relationship between perceived creaminess and tara gum concentration was clearly linear and very strong, with r2=0.95$r^2 = 0.95$. Here is the equation of the least-squares regression line:

$$\text{creaminess}=29.38+101.98\ \text{concentration}$$

$\text{creaminess} = 29.38 + 101.98\ \text{concentration}$

a. What is the slope of the regression line? Is the relationship between creaminess and tara gum concentration positive or negative?
b. What is the predicted perceived creaminess for a tara gum concentration of 0.2 g/100 g? Use r^2 to assess the reliability of this prediction.
c. Using this model, what would be the predicted perceived creaminess for a tara gum concentration of 1 g/100 g? Explain why this is *not* a legitimate prediction.

4.26 Plastic-eating bacteria. Plastics are becoming an extreme source of environmental pollution worldwide. Bioengineers isolated bacteria (*Ideonella sakaiensis 201-F6*) capable of using plastic as a major carbon source. Here is the amount of weight lost over time from a 60-mg piece of plastic placed in a solution containing the new bacteria:[15]

Time (days)	0	7	14	17	21	28	42	56	70
Weight lost (mg)	2	6	11	13	17	25	31	42	52

a. Make a scatterplot of amount of weight lost as a function of time. Describe the relationship. How strong is it?
b. Obtain the equation for the least-squares regression line modeling

weight lost as a function of time. What is the meaning of the slope in the context of this problem? What percent of the variation in weight lost is explained by the time spent in the bacterial solution?

c. Use the regression equation to predict the amount of weight lost after the piece of plastic has stayed in the bacterial solution for 50 days. How much more weight can we expect the piece of plastic to lose after one additional day?

d. Use the regression equation to predict the amount of weight lost after the piece of plastic has stayed in the bacterial solution for 100 days. Does this value make any sense? Explain your answer.

4.27 Grading oysters. Exercises 3.26 and 3.27 (page 82) describe an automated computerized grading system for oysters based on either a 2D or a 3D image-processing approach. The data appear in Table 3.3 (page 86).

a. If you haven’t done so already, make a scatterplot of the relationship between actual volume (response) and 2D volume reconstruction (explanatory), and another scatterplot showing the relationship between actual volume and 3D volume reconstruction. Compare the two image-processing approaches.

b. Obtain the equation for the least-squares regression describing actual volume as a function of 2D reconstruction. Use the regression equation to predict the actual volume of an oyster with a 2D reconstruction of 35 thousand pixels.

c. Obtain the equation for the least-squares regression line describing actual volume as a function of 3D reconstruction. Use the regression equation to predict the actual volume of an oyster with a 3D reconstruction of 4.5 million volume pixels.

d. What percent of the variation in actual oyster volumes is explained by the 2D program? By the 3D program? Which is the more accurate predictive model? Explain your answer.

e. Explain in the context of this exercise why it is important to include both the equation for the least-squares regression line and the value of r2r^2 when reporting on a linear association.

4.28 Keeping water clean. Keeping water supplies clean requires regular measurement of levels of pollutants. The measurements are indirect—a typical analysis involves forming a dye by a chemical reaction with the dissolved pollutant, then passing light through the solution and measuring its “absorbance.” To calibrate such measurements, the laboratory measures known standard solutions and uses regression to relate absorbance and pollutant concentration. This is usually done every day. Here is one series of data on the absorbance for different levels of nitrates. Nitrates are measured in milligrams per liter of water.[16]

Nitrates	50	50	100	200	400	800	1200	1600	2000	2000
Absorbance	7.0	7.5	12.8	24.0	47.0	93.0	138.0	183.0	230.0	226.0

a. Chemical theory says that these data should lie on a straight line. If the correlation is not at least 0.997, something went wrong, and the calibration procedure is repeated. Plot the data and find the correlation. Must the calibration be done again?

b. The calibration process sets nitrate level and measures absorbance. Once established, the linear relationship will be used to estimate the nitrate level in water from a measurement of absorbance. What is the equation of the line used for estimation? What is the estimated nitrate level in a water specimen with absorbance 40?

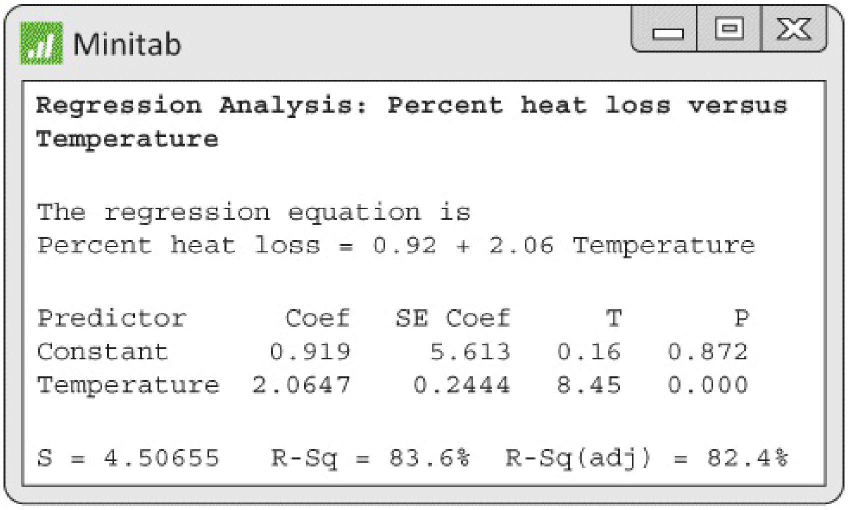
c. Do you expect estimates of nitrate level from absorbance to be highly accurate? Why?

4.29 Predicting tree height. Measuring tree height is not an easy task. How well does trunk diameter predict tree height? A survey of 958 live trees in an old-growth forest in Canada provides us with the following information: The mean tree height is 15.6 meters (m), with standard deviation 13.4 m; the mean diameter, measured at “breast height” (1.3 m above ground), is 23.4 centimeters (cm), and the standard deviation is 23.5 cm. The correlation between the height and the diameter is very high: r=0.96$r = 0.96$.[17]

a. What are the slope and intercept of the regression line to predict tree height from trunk diameter? Draw a graph of this regression line.

b. The tree diameters ranged from 1 to 101 cm. Predict the height of a tree in this area that is 50 cm in diameter at breast height. Use r2r^2 to assess the reliability of this prediction.

4.30 A toucan's beak. Exercise 3.37 (page 91) gives data on toucan beak heat loss, as a percent of total body heat loss from all sources, at various temperatures. The data show that beak heat loss is higher at higher temperatures and that the relationship is roughly linear. Figure 4.9 shows Minitab regression output for these data.



```
Regression Analysis: Percent heat loss versus
Temperature

The regression equation is
Percent heat loss = 0.92 + 2.06 Temperature

Predictor       Coef   SE Coef      T      P
Constant       0.919     5.613   0.16  0.872
Temperature   2.0647    0.2444   8.45  0.000

S = 4.50655   R-Sq = 83.6%   R-Sq(adj) = 82.4%
```

Figure 4.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

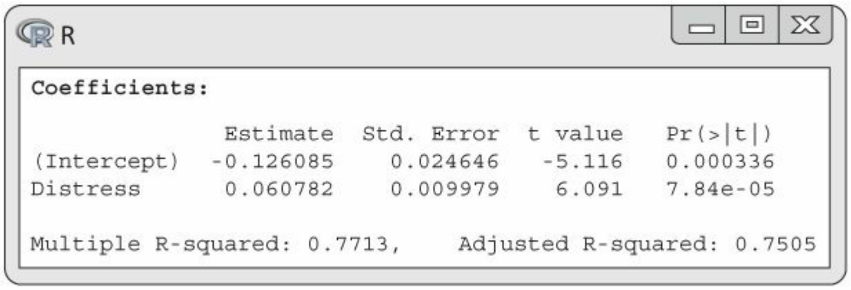
FIGURE 4.9 Minitab regression output for a study of the effect of temperature on toucan beak heat loss, for Exercise 4.30.

a. What is the equation of the least-squares regression line for predicting beak heat loss, as a percent of total body heat loss, from temperature? Use the equation to predict beak heat loss, as a percent of total body heat loss from all sources, at a temperature of 25 degrees Celsius.

b. What percent of the variation in beak heat loss is explained by the straight-line relationship with temperature?

c. Use the information in Figure 4.9 to find the correlation rr between beak heat loss and temperature. How do you know whether the sign of rr is + or −$-$?

4.31 Does social rejection hurt? Exercise 3.36 (page 90) gives data from a study showing that social exclusion causes "real pain." That is, activity in an area of the brain that responds to physical pain increases as distress from social exclusion increases. A scatterplot shows a moderately strong linear relationship. Figure 4.10 shows regression output from software for these data.



```
Coefficients:

              Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  -0.126085    0.024646    -5.116    0.000336
Distress      0.060782    0.009979     6.091    7.84e-05

Multiple R-squared: 0.7713,     Adjusted R-squared: 0.7505
```

Figure 4.10
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 4.10 CrunchIt! regression output for a study of the effects of social rejection on brain activity, for Exercise 4.31.

a. What is the equation of the least-squares regression line for predicting brain activity from social distress score? Use the equation to predict brain activity for a social distress score of 2.0.

b. What percent of the variation in brain activity among these subjects is explained by the straight-line relationship with social distress score?

4.32 Beavers and beetles. Ecologists sometimes find rather strange relationships in our environment. For example, do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so the beetles prefer them. If so, more stumps should produce more beetle larvae. Here are the data:[18]

Stumps	2	2	1	3	3	4	3	1	2	5	1	3
Beetle larvae	10	30	12	24	36	40	43	11	27	56	18	40
Stumps	2	1	2	2	1	1	4	1	2	1	4	
Beetle larvae	25	8	21	14	16	6	54	9	13	14	50	

Analyze these data to see if they support the “beavers benefit beetles” idea. Follow the four-step process (page 58) in reporting your work.

4.33 Always plot your data! Table 4.1 presents four sets of data prepared by the statistician Frank Anscombe to illustrate the dangers of calculating without first plotting the data.[19]

TABLE 4.1 Four data sets for exploring correlation and regression

Data Set A	xx	10	8	13	9	11	14	6	4	12	7	5
	yy	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68
Data Set B	xx	10	8	13	9	11	14	6	4	12	7	5
	yy	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74
Data Set C	xx	10	8	13	9	11	14	6	4	12	7	5
	yy	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73
Data Set D	xx	8	8	8	8	8	8	8	8	8	8	19
	yy	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

a. Without making scatterplots, find the correlation and the least-squares regression line for all four data sets. What do you notice? Use the regression line to predict yy for x=10$x = 10$.

b. Make a scatterplot for each of the data sets and add the regression line to each plot.

c. In which of the four cases would you be willing to use the regression line to describe the dependence of yy on xx? Explain your answer in each case.

4.34 Toxicology of aspartame. The National Toxicology Program evaluates the toxicology of substances suspected of carcinogenicity (causing cancer). One such study evaluated the impact of aspartame added to the diet on the survival rates of mice. For each aspartame concentration studied (in parts per million, ppm), 30 mice were fed and monitored for 40 weeks, and the group’s survival rate was recorded as the percent of mice still alive at the end of the 40 weeks. Here are the results:[20]

Concentration (ppm)	0	3,125	6,250	12,500	25,000	50,000
Survival rate (%)	67	73	57	70	73	60

a. Make a scatterplot of mice survival rate (response) against aspartame concentration (explanatory) and add the regression line. Describe the direction, form, and strength of the relationship. What percent of the variation in mice survival rate can be explained by variations in aspartame concentrations?

b. Make another scatterplot, this time ignoring the last data point (an aspartame concentration of 50,000 ppm). Add the regression line. What happened to the regression line? What does this say about the data point corresponding to the aspartame concentration of 50,000 ppm?

c. Make another scatterplot, this time with all the data points except the one for an aspartame concentration of 6250 ppm. Add the regression line. What happened to the regression line? What does this say about the data point you just removed?

4.35 Drilling into the past. Drilling down beneath a lake in Alaska yields chemical evidence of past changes in climate. Biological silicon, left by the skeletons of single-celled creatures called diatoms, is a measure of the abundance of life in the lake. A rather complex variable based on the ratio of certain isotopes relative to ocean water gives an indirect measure of moisture, mostly from snow. As we drill down, we look further into the past. Here are data from 2300 to 12,000 years ago:[21]

Isotope (%)	**Silicon (mg/g)**	**Isotope (%)**	**Silicon (mg/g)**	**Isotope (%)**	**Silicon (mg/g)**
−–19.90	97	−–20.71	154	−–21.63	224
−–19.84	106	−–20.80	265	−–21.63	237
−–19.46	118	−–20.86	267	−–21.19	188
−–20.20	141	−–21.28	296	−–19.37	337

a. Make a scatterplot of silicon (response) against isotope (explanatory). Ignoring the outlier, describe the direction, form, and strength of the relationship. The researchers say that this relationship and relationships among other variables they measured are evidence for cyclic changes in climate that are linked to changes in the sun’s activity.

b. The researchers single out one point: “The open circle in the plot is an outlier that was excluded in the correlation analysis.” Circle this outlier on your graph. What is the correlation with and without this point? The point strongly influences the correlation. Explain why the outlier moves r in the direction revealed by your calculations.

c. Is the outlier also strongly influential for the regression line? Calculate and draw on your graph the regression lines with and without the outlier, then discuss what you see. Explain why adding the outlier moves the regression line in the direction shown on your graph.

4.36 Climate change and ecosystems. Many studies have documented a clear pattern of climate change toward warmer temperatures globally. What has been the subject of controversy, however, is the role of human activity and the impact of climate change on ecosystems. A survey of the North Sea over a 25-year period (1977 to 2001) shows that, as winter sea temperatures have warmed up, fish populations have moved farther north. Here are the data for mean winter sea-bottom temperature (in degrees Celsius) and mean northern latitude (in degrees) of anglerfish populations each year over the 25-year study period:[22]

Temperature (o◦ C)	**Latitude (o◦)**	**Temperature (o◦ C)**	**Latitude (o◦)**	**Temperature (o◦ C)**	**Latitude (o◦)**
6.26	57.20	6.52	57.72	7.09	58.07
6.26	57.96	6.68	57.83	7.13	58.49
6.27	57.65	6.76	57.87	7.15	58.28
6.31	57.59	6.78	57.48	7.29	58.49
6.34	58.01	6.89	58.13	7.34	58.01
6.32	59.06	6.90	58.52	7.57	58.57
6.37	56.85	6.93	58.48	7.65	58.90
6.39	56.87	6.98	57.89		
6.42	57.43	7.02	58.71		

a. Make a scatterplot of the relationship between latitude of anglerfish populations (yy) and temperature (xx). Describe the form and strength of that relationship.

b. Give the equation for the regression line to predict latitude from temperature and the value of the correlation. Add the regression line to your scatterplot.

c. The scatterplot shows one outlier of the relationship. Remove that point from the data set and calculate the new regression line and correlation. Add the new regression line to your original scatterplot, making sure to use a different color or line thickness.

d. How do the regression line and correlation omitting the outlier compare with the original ones found in part b? Use your findings to discuss how influential the outlier is.

4.37 Predicting seed count. Exercise 3.32 (page 88) examined the relationship between mean seed weight (in milligrams, mg)and mean number of seeds produced in a year by several common tree species. After converting the data from Table 3.5 (page 89)into a logarithm scale, a scatterplot of the converted data shows a linear relationship that can be described by the least-square regression equation

y^=4.238-0.567x

$$\hat{y} = 4.238 - 0.567x$$

where y^$\hat{y}$ is the log of seed count and xx is the log of seed weight (both in base 10).

a. Using this equation, predict the log of seed count y^$\hat{y}$ for a tree species with a mean seed weight of 1000 mg (x=3$x = 3$).

b. Following Example 4.2, undo the logarithm transformation to find the predicted mean annual seed count for a tree species with a mean seed weight of 1000 mg.

4.38 Counting carnivores. Ecologists look at data to learn about nature’s patterns. One pattern they have found relates the size of a carnivore (body mass in kilograms) to how many of those carnivores reside in an area. One way to measure “how many” is to count carnivores per 10,000 kilograms of their prey in the area. Table 4.2 gives average data for 25 carnivore species.[23]

TABLE 4.2 **Size and abundance of carnivores**

Carnivore species	Body mass (kg)	Abundance	Carnivore species	Body mass (kg)	Abundance
Least weasel	0.14	1656.49	Eurasian lynx	20.0	0.46
Ermine	0.16	406.66	Wild dog	25.0	1.61
Small Indian mongoose	0.55	514.84	Dhole	25.0	0.81
Pine marten	1.3	31.84	Snow leopard	40.0	1.89
Kit fox	2.02	15.96	Wolf	46.0	0.62
Channel Island fox	2.16	145.94	Leopard	46.5	6.17
Arctic fox	3.19	21.63	Cheetah	50.0	2.29
Red fox	4.6	32.21	Puma	51.9	0.94
Bobcat	10.0	9.75	Spotted hyena	58.6	0.68
Canadian lynx	11.2	4.79	Lion	142.0	3.40
European badger	13.0	7.35	Tiger	181.0	0.33
Coyote	13.0	11.65	Polar bear	310.0	0.60
Ethiopian wolf	14.5	2.70			

a. Plot and describe the relationship between carnivore abundance (response variable) and body mass (explanatory variable). Would it be appropriate to conduct a linear regression analysis for the data as is? Explain your answer.

b. When a relationship is clearly not linear, transforming the data can sometimes help. For each species, compute the logarithm of body mass and the logarithm of abundance. Plot and describe the relationship between the log of abundance and the log of body mass. Is linear regression appropriate for the transformed data?

c. The least-squares regression for the transformed data gives the equation

$$\log(\text{abundance})=1.95-[1.05\times\log(\text{body mass})]$$

$\log(\text{abundance}) = 1.95 - [1.05 \times \log(\text{body mass})]$

Predict the abundance of a carnivore species whose individuals weigh about 1 kg. Make sure to transform the units both ways (remember that y=10log(y)$y = 10^{\log(y)}$).

4.39 Are big hospitals bad for you? A study shows that there is a positive correlation between the size of a hospital (measured by its number of beds xx) and the median number of days yy that patients remain in the hospital. Does this mean that you can shorten a hospital stay by choosing a small hospital?Why?

4.40 How's your self-esteem? People who do well in life tend to feel good about themselves. Perhaps helping students feel good about themselves will help them do better both in school and later life. Raising self-esteem became for a time a goal in many schools. California even created a state commission to advance the cause, hoping that it would help reduce crime, drug use, and teen pregnancy rates (the commission was ended in 1995). Can you think of explanations for the association between high self-esteem and good school performance other than "Self-esteem causes better work in school"?

4.41 Another cup of coffee? Researchers assessed coffee drinking every four years (using validated food questionnaires) for more than 200,000 adults followed for as long as 30 years. They found that higher coffee consumption was associated with lower risk of total mortality. The published report states, "Compared with nondrinkers, coffee consumption of 1 to 5 cups per day was associated with lower risk of mortality, whereas coffee consumption of more than 5 cups per day was not associated with risk of mortality."[24]

a. Should doctors encourage adult patients to drink more coffee based on this study? Explain your reasoning.
b. Can you think of any reason why adults who do not drink coffee may have a lower longevity than adults who do?

4.42 Will smoking kill you? A congressman once posted on his website that smoking doesn't kill because two out of every three smokers do not die from a smoking-related illness.[25] Must an association be deterministic without any variability to be able to legitimately conclude causality? What lurking variables may influence a smoker's chance of dying from a smoking-related illness?

4.43 HPV and cervical cancer. The human papillomavirus (HPV) has been determined to cause cervical cancer. The evidence for this conclusion is available online. Typing DOI:10.1136/jcp.55.4.244 in a search engine will give you a link to the scientific article titled “The causal relation between human papillomavirus and cervical cancer.” The article is long but the key points are nicely summarized in a series of small gray boxes with a title containing the word “Evaluation.” Match the gray boxes in the article to the five criteria we listed for establishing causation in the absence of experiments. Are these five criteria met in the context of HPV infection and cervical cancer?

4.44 Influence in regression. The *Correlation and Regression* applet allows you to animate Figure 4.4. Click to create a group of 10 points in the lower-left corner of the scatterplot with a strong straight-line pattern (correlation approximately 0.9). Click the “Show least-squares line” box to display the regression line.

a. Add one point at the upper right that is far from the other 10 points but exactly on the regression line. Why does this outlier have no effect on the line even though it changes the correlation?
b. Now use the mouse to drag this last point straight down. You see that one end of the least-squares line chases this single point, while the other end remains near the middle of the original group of 10 points. What makes the last point so influential?

4.45 Is regression useful? In Exercise 3.39 (page 91) you used the *Correlation and Regression* applet to create three scatterplots having correlation approximately r=0.7$r = 0.7$ between the horizontal variable xx and the vertical variable yy. Create three similar scatterplots again and click the “Show least-squares line” box to display the regression lines. Correlation r=0.7$r = 0.7$ is considered reasonably strong in many areas of work. Because there is a reasonably strong correlation, we might use a regression line to predict yy from xx. In which of your three scatterplots does it make sense to use a straight line for prediction?

Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Short answers are not available for these exercises to enable their use for class work or assignments.*

4.46 Children's growth. The *Large.FEV* data file contains a number of morphological and physiological measurements for a large sample of children.[26] We can use the data to study growth patterns in children.

a. Plot the relationship between age in years (explanatory) and height in inches (response) in boys and girls in two separate graphs. Compare both graphs. How do boys and girls differ? In what way are their growth patterns similar?
b. The nonlinear pattern of growth over the whole age range suggests that we should study younger children (ages 10 and younger) separately from adolescents (ages 11 and older). Plot the relationship between age and height for the younger children and obtain the least-squares regression equation. What percent of the variation in height is explained by this model?
c. We noticed in the first analysis in part a that boys and girls start to differ notably around the time of puberty. Plot the relationship between age and height for adolescent girls and obtain the least-squares regression equation. What percent of the variation in height is explained by this model? Do the same for adolescent boys. Describe the main differences between the growth patterns of adolescent boys and girls.

CHAPTER 5 Two-Way Tables

Wendy Shattil and Bob Rozinski/Getty Images

IN THIS CHAPTER WE COVER...

- Marginal distributions
- Conditional distributions
- Simpson’s paradox

We have concentrated on relationships in which at least the response variable is quantitative. Now we will describe relationships between two categorical variables. Some variables—such as sex, species, and color—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes, for example, age groups. Published data often appear in grouped form to save space. To analyze categorical data, we use the *counts* or *percents* of individuals that fall into various categories.

EXAMPLE 5.1 Foot health

Foot pain is a very common musculoskeletal complaint in the United States, especially among older adults. Foot pain is more frequent among women than among men, possibly because men and women wear different types of shoes over their lifetimes.

A research group surveyed individuals from a large established cohort of residents from Framingham, Massachusetts, a group now made up of mainly older individuals (65 years of age, on average). The participants were asked to select from a list which type of footwear they had worn most regularly in the past. The answers were then categorized as shoes providing good, average, or poor structural foot support. Table 5.1 presents the results for the men and the women in the study.[1]

TABLE 5.1 Study participants by shoe type and gender

	Gender		
Shoe type	**Men**	**Women**	**Total**
Good support	94	137	231
Average support	1348	581	1929
Poor support	30	1182	1212
Total	1472	1900	3372

Both variables, shoe type and gender, are categorical. Table 5.1 is a **two-way table** because it describes the relationship between these two categorical variables. Shoe type is the **row variable** because each row in the table describes one type of shoe classified in the study by the structural support it provides. Gender is the **column variable** because each column describes either men or women. The entries in the table are the counts of study participants in each support-by-gender class.

MARGINAL DISTRIBUTIONS

How can we best grasp the information contained in Table 5.1? First, *look at the distribution of each variable separately*. The distribution of a categorical variable says how often each outcome occurred. The “Total” column at the right of the table contains the totals for each of the rows. These row totals give the distribution of shoe types (the row variable) in the study: 231 participants had worn mainly shoes providing good support; 1929, average support; and 1212, poor support. In the same way, the “Total” row at the bottom of the table gives the gender distribution: The study interviewed 1472 men and 1900 women.

If the row and column totals are missing, the first thing to do in studying a two-way table is to calculate them. The distributions of shoe type alone and gender alone are called **marginal distributions** because they appear at the right and bottom margins of the two-way table, respectively.

Percents (proportions) are often more informative than counts. We can display the marginal distribution of shoe types in terms of percents by dividing each row total by the table total and converting the result to a percent.

EXAMPLE 5.2 Calculating a marginal distribution

The percent of study participants who wore mostly shoes providing good support is

good support totaltable total=2313372=0.0685=6.9%

$$\frac{\text{good support total}}{\text{table total}}=\frac{231}{3372}=0.0685=6.9\%$$

Do two more such calculations to obtain the marginal distribution of shoe types in percents. Here is the complete distribution:

Shoe type	Percent
Good support	2313372=6.9%
Average	19293372=57.2% $\frac{1929}{3372}=57.2\%$

support	
Poor support	12123372=35.9%$\frac{1212}{3372} = 35.9\%$

It seems that only a small percent of individuals in the study tended to wear mainly shoes providing good support. The total sums to 100% because everyone in the study selected one type of shoe as the one they tended to wear most often.

Sometimes percents add up to slightly less or more than 100% because we round the computed values. This discrepancy is called **roundoff error.**

Each marginal distribution from a two-way table is a distribution for a single categorical variable. As we saw in Chapter 1, we can use a bar graph or a pie chart to display such a distribution. Figure 5.1 is a bar graph of the distribution of shoe types for all the study participants.

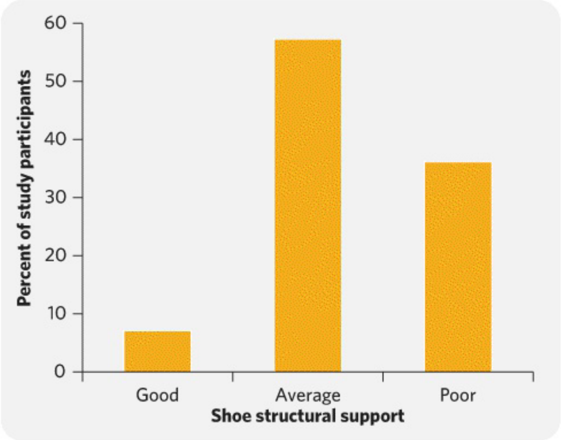


Figure 5.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 5.1 A bar graph of the distribution of shoes worn by the type of structural support they offer. This is one of the marginal distributions for Table 5.1.

In working with two-way tables, you must calculate lots of percents. Here's a tip to help decide which fraction gives the percent you want. Ask, "Which group represents the total that I want a percent of?" The count for that group is the denominator of the fraction that leads to the percent. In Example 5.2, we want a percent of "study participants," so the count of study participants (the table total) is the denominator.

APPLY YOUR KNOWLEDGE

5.1 Peanut allergy. Peanut allergies are becoming increasingly common in Western countries. Some evidence points to the timing of peanuts' first introduction in the diet as an influential factor, raising the question of whether pediatricians should recommend early exposure or avoidance. A study enrolled infants with a diagnosed peanut allergy and randomly assigned them to either completely avoid peanuts or consume peanuts in small amounts regularly until they reached 60 months of age. At the end of the study, 18 of the 51 infants who had avoided peanuts were still allergic to peanuts. In contrast, 5 of the 47 infants who had consumed peanuts were still allergic to peanuts.[2]

a. What are the two variables described? Organize the study findings into a two-way table of counts similar to Table 5.1.

b. How many infants do these data represent? How many of them were assigned to avoid peanuts during the study?

c. Give the marginal distribution of peanut allergy at 60 months of age, both as counts and as percents.

5.2 Altruism in prairie dogs. Prairie dogs are social rodents with sophisticated calls capable of identifying various kinds of predators to alert their kin. The warning calls are reminiscent of the barks of dogs —hence the species' name. Raising the alarm when a predator is spotted is the group's only defense mechanism, but it places the caller at a higher risk of predation. A study examined whether proximity to a predator (a stuffed badger controlled remotely by an experimenter) influences the likelihood that a prairie dog will raise the alarm. The following table shows the count of prairie dogs raising the alarm or not when they were either near the predator or far from it:[3]

	Alarm call	**No call**
Near	29	80
Far	55	78

a. How many prairie dogs were monitored in this experiment?

b. How many of these prairie dogs were near the predator?

c. Give the marginal distribution of reaction (alarm call or no call) as percents and display them in a bar graph.

Wendy Shattil and Bob Rozinski/Getty Images

5.3 Marginal distributions aren't the whole story. Here are the row and column totals for a two-way table with two rows and two columns:

aa	bb	50
cc	dd	50
60	40	100

Find *two different* sets of counts a,$a,$ b,$b,$ c,$c,$ and dd for the body of the table that give these same totals. This shows that the relationship between two variables cannot be obtained from the two individual distributions of the variables.

CONDITIONAL DISTRIBUTIONS

Table 5.1 contains much more information than just the two marginal distributions of shoe type alone and gender alone. *Marginal distributions tell us nothing about the relationship between two variables.* To describe a relationship between two categorical variables, we must calculate some well-chosen percents from the counts given in the body of the table. We use percents because counts are often hard to compare. For example, there are 1900 women in the foot health study but only 1472 men, so comparing counts for men and women would obviously be misleading.

We may want to compare the types of shoes worn by men and women. To do this, we compare percents for women alone with percents for men alone. First, to study the type of shoes that men say they tend to wear most often, we look only at the "Men" column in Table 5.2. To find the percent *of men* who wore mostly shoes providing good support, divide the count of such men by the total number of men (the column total):

TABLE 5.2 Shoe type by gender: men's group

	Gender		
Shoe type	**Men**	**Women**	**Total**
Good support	94	137	231
Average support	1348	581	1929
Poor support	30	1182	1212
Total	1472	1900	3372

men wearing shoes with good supportmen's column total=941472=0.064=6.4%

$$\frac{\text{men wearing shoes with good support}}{\text{men's column total}} = \frac{94}{1472} = 0.064 = 6.4\%$$

Doing this for all three entries in the "Men" column gives the *conditional distribution* of shoe types among men. We use the term "conditional" because this distribution describes only study participants who satisfy the chosen condition—in this case, being male.

MARGINAL AND CONDITIONAL DISTRIBUTIONS

In a two-way table of counts, the **marginal distribution** of one of the categorical variables is the distribution of values of that variable among all individuals described by the table.

A **conditional distribution** of a variable is the distribution of values of that variable among only individuals who have a given value of the other variable. There is a separate conditional distribution for each value of the other variable.

EXAMPLE 5.3 Comparing men and women: conditional distribution of shoe type, given a particular gender

STATE: Foot pain is a common ailment in older individuals, especially among women. Poor footwear is thought to be a contributing factor of foot pain. We want to know how older men and women differ in the type of shoes they wore most regularly in the past.

PLAN: Make a two-way table of shoe type by gender. Find the conditional distributions of shoe type for men alone and for women alone. Compare these two distributions.

SOLVE: Comparing conditional distributions reveals the nature of the association between shoe type and gender. We use the data displayed in Table 5.1. Look first at just the "Men" column to find the conditional distribution for men, then at just the "Women" column to find the conditional distribution for women. Here are the calculations and the two conditional distributions:

Shoe type	Men	Women
Good support	941472=6.4% $\frac{94}{1472} = 6.4\%$	1371900=7.2% $\frac{137}{1900} = 7.2\%$
Average	13481472=91.6% $\frac{1348}{1472} = 91.6\%$	5811900=30.6% $\frac{581}{1900} = 30.6\%$

support		
Poor support	301472=2.0% $\frac{30}{1472} = 2.0\%$	11821900=62.2% $\frac{1182}{1900} = 62.2\%$

Each set of percents within a gender condition adds to 100% because everyone within the gender group selected one of the three shoe types. The bar graph in Figure 5.2(a) compares the percents of shoe types (offering good, average, and poor structural support) among men and among women in the study. Figure 5.2(b) displays the same percents in a “stacked” bar graph, which helps us visualize the fact that the conditional distribution within each gender adds up to 100% (the stacked bar graph format is closely related to that of the pie chart).

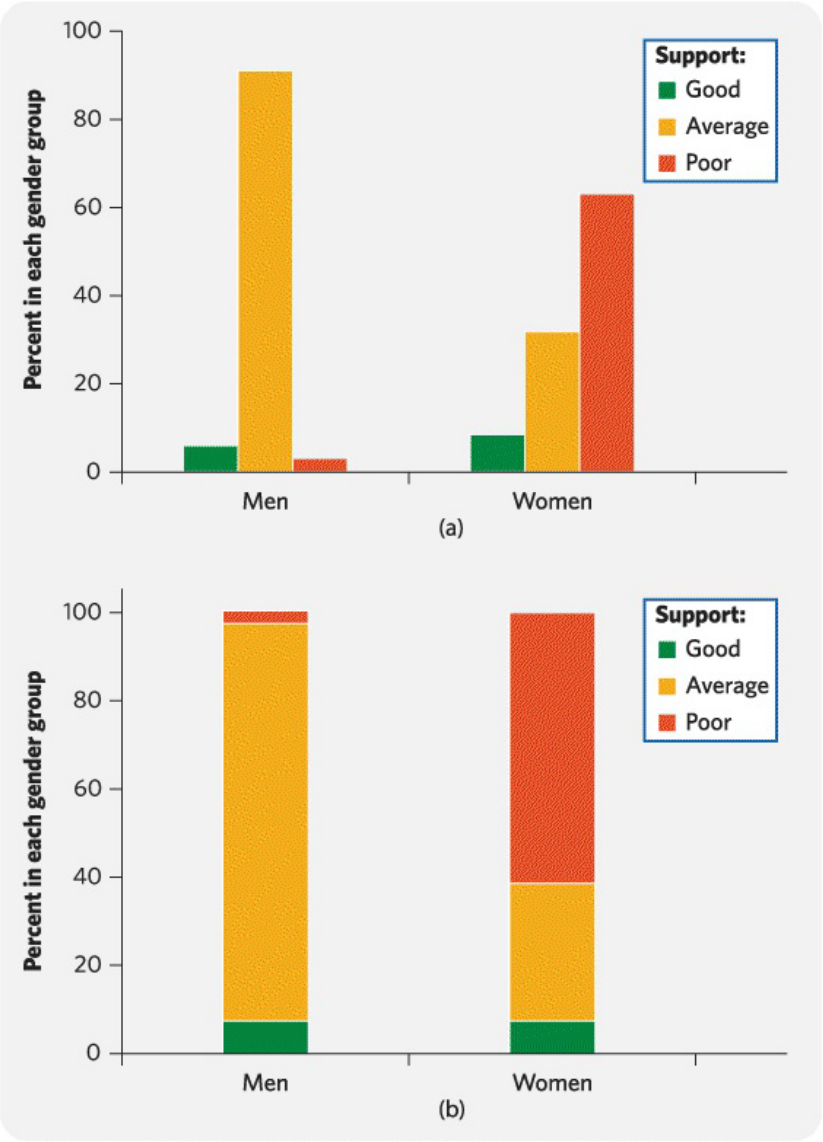


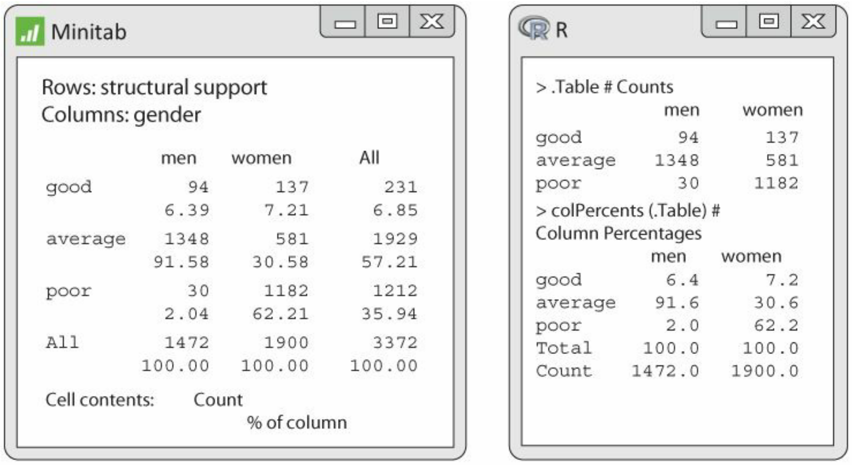
Figure 5.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 5.2 (a) Bar graph comparing the percents of shoe types (offering good, average, and poor structural support) among men and among women in the study. Men and women wore all three types of shoes, but the percent of shoes offering poor structural support was higher among the women. (b) The same percents are displayed in a “stacked” bar graph, which helps visualize the fact that the conditional distribution within each gender adds up to 100%.

Fabrice LEROUGE/Getty Images

CONCLUDE: Men overwhelmingly wore shoes providing average foot support. In contrast, women overwhelmingly wore shoes with either average or poor foot support. In particular, women were much more likely than men to have worn shoes providing poor support (62.2% compared with only 2.0%). This difference might explain why foot pain is more common among women than among men, although we should be careful not to assign causality to an observed association.

Software will perform these calculations for you once you understand how to enter and label the data for your particular program. Most programs allow you to choose which conditional distributions you want to compare. The outputs from Minitab and R in Figure 5.3 compare the conditional distributions of shoe type, given gender. These percents agree with the results in Example 5.3. Minitab also provides the marginal distribution of shoe type for all study participants, matching the values we computed in Example 5.2.



```
Rows: structural support
Columns: gender

             men    women      All
good          94      137      231
            6.39     7.21     6.85
average     1348      581     1929
           91.58    30.58    57.21
poor          30     1182     1212
            2.04    62.21    35.94
All         1472     1900     3372
          100.00   100.00   100.00

Cell contents:   Count
                      % of column
```

```
> .Table # Counts
          men   women
good       94     137
average  1348     581
poor       30    1182
> colPercents (.Table) #
Column Percentages
          men    women
good       6.4     7.2
average   91.6    30.6
poor       2.0    62.2
Total    100.0   100.0
Count   1472.0  1900.0
```

Figure 5.3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 5.3 Minitab and R outputs of the two-way table of shoe support types for men and women with each entry as a percent of its column total. The percents in each column give the conditional distribution of shoe support type for one gender. The "All" column in the Minitab output also gives the marginal distribution of shoe support type for the whole study (both genders combined).

Recall that there are two sets of conditional distributions for any two-way table. Example 5.3 looked at the conditional distributions of shoe type for each gender. Alternatively, we could examine the conditional distributions of gender for the three shoe types. Starting with the shoe type providing good support, we find the conditional distribution of gender by looking only at the first row of the two-way table, as highlighted in Table 5.3. We then proceed to the other two conditional distributions by focusing, in turn, on the second and third rows. Here are all the computations and results:

Shoe type	Men	Women
Good support	94231=40.7%$\frac{94}{231} = 40.7\%$	137231=59.3%$\frac{137}{231} = 59.3\%$
Average support	13481929=69.9%$\frac{1348}{1929} = 69.9\%$	5811929=30.1%$\frac{581}{1929} = 30.1\%$
Poor support	301212=2.5%$\frac{30}{1212} = 2.5\%$	11821212=97.5%$\frac{1182}{1212} = 97.5\%$

TABLE 5.3 Gender by shoe type: good support group

	Gender		
Shoe type	**Men**	**Women**	**Total**
Good support	94	137	231
Average support	1348	581	1929
Poor support	30	1182	1212
Total	1472	1900	3372

Each row adds to 100% because each conditional distribution includes all study participants (both men and women) who wore a particular shoe type. The most striking finding here is that the study participants who wore mostly shoes providing poor foot support are overwhelmingly women (97.5% women and only 2.5% men).

No single graph or numerical summary can portray the form of the relationship between categorical variables. Instead, you must think about which comparisons you want to display graphically and which percents to compute. Here is a hint to help you decide: *If there is an explanatory–response relationship, compare the conditional distributions of the response variable for the separate values of the explanatory variable.* In the foot health study, we know that men and women make different footwear choices, so gender would be our explanatory variable. Thus, the most interesting comparison is the one made in Example 5.3: comparing the conditional distributions of shoe type for men and for women.

APPLY YOUR KNOWLEDGE

5.4 Foot health: current footwear. The study in Example 5.1 also asked participants to describe their current footwear. Here are the findings:

Shoe type	Men	Women
Good support	612	785
Average support	836	862
Poor support	24	253

How do older men and women from the Framingham cohort differ in the type of shoes they currently wear most regularly? Follow the four-step process, as illustrated in Example 5.3.

5.5 Altruism in prairie dogs, continued. The two-way table in Exercise 5.2 describes a study of warning-call behavior in prairie dogs exposed to different predation risks as defined by the level of proximity to a predator.

a. Predator proximity is the explanatory variable in this study. Find the conditional distribution of call behavior, given predator proximity.

b. Based on your calculations, describe the differences in call behavior when a predator is near or far. Use both a graph and a written explanation.

5.6 Peanut allergy, continued. Exercise 5.1 describes a study of two long-term approaches to handling peanut allergy in infants. Do the data indicate that one approach is more beneficial? Follow the four-step process, as illustrated in Example 5.3.

SIMPSON’S PARADOX

As is the case with quantitative variables, the effects of lurking variables can change or even reverse relationships between two categorical variables. Here is an example that demonstrates the surprises that can await the unsuspecting user of data.

STATS IN YOUR WORLD

DO LEFT-HANDERS DIE EARLY?

Yes, said a study of 1000 deaths in California. Left-handed people died at an average age of 66 years; right-handers, at 75 years of age. Should left-handed people fear an early death? No—the lurking variable has struck again, resulting in Simpson’s paradox. Older people grew up in an era when many natural left-handers were forced to use their right hands. As a consequence, right-handers are more common among older people, and left-handers are more common among the young. When we look at deaths, the left-handers who die are younger, on average, because left-handers in general are younger. Mystery solved.

EXAMPLE 5.4 Death rates vary by age group

When comparing death rates among subgroups, it is important to consider that these groups may have differing age distributions. The Centers for Disease Control and Prevention (CDC) provides a *theoretical* example designed to make the computations obvious.[4] Consider two communities, A and B, with the following data:

	Community A			**Community B**		
Age group	**Deaths**	**Population**	**Death rate per 1,000**	**Deaths**	**Population**	**Death rate per 1,000**
0 to 34	20	1,000	20	180	6,000	30
35 to 64	120	3,000	40	150	3,000	50

65 and older	360	6,000	60	70	1,000	70
Total	500	10,000	50	400	10,000	40

From the last row, we see that, overall, community A has a higher death rate (50 per 1000) than community B (40 per 1000). Yet, among individuals age 0 to 34, the death rate is lower in community A (20 per 1000) than in community B (30 per 1000). The same holds true for individuals age 35 to 64 (40 per 1000 versus 50 per 1000) and for individuals age 65 and older (60 per 1000 versus 70 per 1000). We reach completely opposite conclusions when we simply compare the two communities and when we compare them within each age group. How can this be?

Notice how, in each community, the death rate is—unsurprisingly—higher for the older age groups. Now notice how different the two communities in this fictitious example are in terms of age distribution: Community A has an older population (more than half of the individuals are 65 years and older), whereas community B has a younger population (more than half of the individuals are 0 to 34 years old). Because death rates vary by age group, and because the age distributions are markedly different in these two communities, comparing communities A and B regardless of age would be like comparing apples and oranges. It would not be a fair, equivalent comparison.

Without paying attention to age, community A appears to have a higher mortality rate. Looking at the data one age group at a time brings paradoxically a completely different conclusion: Within each specific age group, community B is the one with the higher mortality rate. This difference in conclusions is called *Simpson's paradox*. The explanation for this paradox is that age distribution is a *lurking variable* here, and ignoring it would lead us to the wrong conclusion.

SIMPSON'S PARADOX

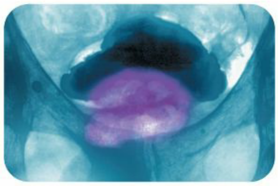
An association or comparison that holds for all of several groups can reverse direction when the data are combined to form a single group. This reversal is called **Simpson's paradox.**

The lurking variable in Simpson's paradox is categorical. That is, it breaks the individuals into groups, such as age groups. Simpson's paradox is just an extreme form of the fact that observed associations can be misleading when there are lurking variables. Here is a real-life example provided by the New York State Department of Health specifically to illustrate the importance of taking age groups into account when comparing mortality rates across different groups.

EXAMPLE 5.5 Prostate cancer, race, and age

In upstate New York (New York state excluding New York City), 121 black men and 1359 white men died from prostate cancer in 1994.[5] Because we know how many individuals lived in upstate New York that year, we can find the prostate cancer mortality rates in these two groups:

Black men:	28.9 deaths per 100,000 men
White men:	28.7 deaths per 100,000 men



James Cavallini/Science Source

The conditional distribution of prostate cancer mortality, given a man's race, shows that black men and white men were, overall, about equally likely to die from prostate cancer in upstate New York in 1994. We should consider, however, that prostate cancer is more common among older men than among younger men. Here are the prostate cancer mortality rates broken down by both race and age group:

Black men younger than 65 years:	4.5 deaths per 100,000 men

White men younger than 65 years:	1.2 deaths per 100,000 men
Black men 65 years and older:	462.1 deaths per 100,000 men
White men 65 years and older:	228.8 deaths per 100,000 men

We now see that in each age group, black men were more than twice as likely to die from prostate cancer than white men. This is a very different conclusion from the one we would reach if we ignored age distribution as a lurking variable.

The New York State Department of Health offers this example to illustrate why overall mortality rates often need to be presented in an *age-adjusted* form. This epidemiological approach is commonly used when comparing different groups if age distribution is a lurking variable. The details of how to compute age-adjusted population rates is beyond the scope of an introductory statistics textbook. The objective of the adjustments, however, is simply to avoid reaching incorrect conclusions because of Simpson’s paradox.

Similarly, many medical studies examine categorical variables such as “dead/still alive” or “condition improved/not improved.” Because pooling data from heterogeneous groups can lead to Simpson’s paradox, data gathered in different studies are not pooled together even if the studies looked at the same set of variables. Instead, a more complex method called *meta-analysis*, which compares the conclusions reached by each study, has been developed and is often found in the biomedical literature. This method is also beyond the scope of our book.

Lastly, recall that association does not by itself imply causation. We discussed this issue at length in Chapter 4 in the context of the association between two quantitative variables. The warning is just as valid when the variables are categorical.

APPLY YOUR KNOWLEDGE

5.7 Kidney stones. A study compared the success rates of two procedures for removing kidney stones: open surgery and percutaneous nephrolithotomy (PCNL), a minimally invasive technique. Here are the number of procedures that were successful or unsuccessful in removing patients' kidney stones, by type of procedure. A separate table is given for patients with small kidney stones and for patients with large stones:[6]

Small stones

	Open surgery	PCNL
Success	81	234
Failure	6	36

Large stones

	Open surgery	PCNL
Success	192	55
Failure	71	25

a. Find the percent of kidney stones, combining the data for small and large stones, that were successfully removed for each of the two medical procedures. Which procedure had the higher overall success rate?

b. What percent of all small kidney stones were successfully removed? What percent of all large kidney stones were successfully removed? Which type of kidney stone appears to be easier to treat?

c. Now find the percent of successful procedures of each type for small kidney stones only. Do the same for large kidney stones. PCNL led to worse outcomes for *both* small and large kidney stones, yet it had a better overall success rate. That sounds impossible. Explain carefully, referring to the data, how this paradox can happen.

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

5.8 Smoking and staying alive. In the mid-1970s, a medical study contacted randomly chosen people in a district in England. Here are data on the 1314 women contacted who either were current smokers or had never smoked. The tables classify these women by their smoking status and age at the time of the survey and whether they were still alive 20 years later.[7]

Ages 18 to 44

	Smoker	**Not**
Dead	19	13
Alive	269	327

Ages 45 to 64

	Smoker	**Not**
Dead	78	52
Alive	169	147

Ages 65+

	Smoker	**Not**
Dead	42	165
Alive	7	28

a. Using these data, create a two-way table of smoking (yes or no) by dead or alive. What percent of the smokers stayed alive for 20 years? What percent of the nonsmokers survived? It seems surprising that a higher percent of the smokers stayed alive.

b. The age of the women at the time of the study is a lurking variable. Show that within each of the three age groups in the data, a higher percent of nonsmokers remained alive 20 years later. Explain why this is an example of Simpson's paradox.

c. The study authors give this explanation: "Few of the older women (over 65 at the original survey) were smokers, but many of them had died by the time of follow-up.” Compare the percents of smokers in the three age groups to verify the explanation.

CHAPTER 5 SUMMARY

- A **two-way table** of counts organizes data about two categorical variables. Levels of the **row variable** are used as labels for the rows that run across the table, and levels of the **column variable** are used as labels for the columns that run down the table. Two-way tables are often used to summarize large amounts of information by grouping outcomes into categories.
- The **row totals** and **column totals** in a two-way table give the **marginal distributions** of the two individual variables. It is clearer to present these distributions as percents of the table total. Marginal distributions tell us nothing about the relationship between the variables.
- There are two sets of **conditional distributions** for a two-way table: the distributions of the row variable for each level of the column variable, and the distributions of the column variable for each level of the row variable. Comparing one set of conditional distributions is one way to describe the association between the row and the column variables.
- To find the **conditional distribution** of the row variable for one specific level of the column variable, look only at that one column in the table. Find each entry in the column as a percent of the column total.
- **Bar graphs** are a flexible means of presenting categorical data. There is no single best way to describe an association between two categorical variables, unless one of them is clearly the explanatory variable.
- A comparison between two variables that holds for each individual value of a third variable can be changed or even reversed when the data for all values of the third variable are combined. This reversal is called **Simpson's paradox.** Simpson's paradox is an example of the effect of lurking variables on an observed association.

THIS CHAPTER IN CONTEXT

In Chapters 3 and 4 we considered relationships between two quantitative variables. In this chapter we use two-way tables to describe relationships between two *categorical* variables. To explore relationships between two categorical variables, we examine their conditional distributions. The conditional distribution of one of the categorical variables is the distribution of that variable among only individuals who have a given value of the other variable.

We will revisit conditional distributions in Chapter 10 in the context of probability. There, we will see that the probability of an event can be computed given some known information, resulting in the concept of conditional probability. In fact, two-way tables are a useful tool to represent probabilities and help compute conditional probabilities.

When exploring relationships between two categorical variables, changes in the pattern of the conditional distribution of one variable as the value of the other varies provide information about the nature of the relationship. When this pattern does not change, it suggests that there is no relationship between the variables. In Chapters 20 and 22, we will use statistical inference to test whether the observed relationship can be explained by chance alone or whether we can infer that the relationship is valid in some broader sense.

As was the case in Chapters 3 and 4, we must be careful not to assume that an observed association implies a cause-and-effect relation. A lurking variable may influence both variables studied and either create the appearance of an association or change the nature of the association. Simpson’s paradox is an example of how such a lurking variable can mislead us. We will revisit the concept of lurking variables in Chapters 6 and 7 in the context of data collection.

CHECK YOUR SKILLS

A study was designed to assess the effect of echinacea extracts on the rates of infection of individuals voluntarily exposed to the common-cold rhinovirus. Some subjects took an echinacea extract daily, starting 7 days before viral exposure and continuing until the end of the study. Another group of subjects received the extract from the time of viral exposure onward only. A third and last group received just an inactive solution ("placebo group"). Following is a two-way table of the number of diagnosed clinical colds among infected subjects in each treatment group.[8] *Exercises 5.9 to 5.15 are based on this table.*

	Cold	**No cold**
7 days before viral exposition and onward	73	59
From time of viral exposition onward	88	43
Placebo (no echinacea)	58	30

5.9 How many individuals are described by this table?

a. 219
b. 351
c. Need more information

5.10 What percent of infected subjects were diagnosed with a cold?

a. Approximately 22%
b. Approximately 38%
c. Approximately 62%

5.11 Your percent from Exercise 5.10 is part of

a. the marginal distribution of outcome (cold or no cold diagnosed).
b. the marginal distribution of treatment.
c. the conditional distribution of outcome, given treatment.

5.12 What percent of the infected subjects in the placebo group were diagnosed with a cold?

a. Approximately 58%
b. Approximately 62%
c. Approximately 66%

5.13 Your percent from Exercise 5.12 is part of

a. the marginal distribution of outcome.
b. the conditional distribution of outcome, given treatment.
c. the conditional distribution of treatment, given outcome.

5.14 What percent of those diagnosed with a cold were in the placebo group?

a. Approximately 26%
b. Approximately 58%
c. Approximately 66%

5.15 Your percent from Exercise 5.14 is part of

a. the marginal distribution of outcome.
b. the conditional distribution of outcome, given treatment.
c. the conditional distribution of treatment, given outcome.

In 2015, Gallup conducted a survey of coffee consumption using a random sample of 1009 American adults. Here is a quote from the resulting online report: "Sixty-four percent of U.S. adults report drinking at least one cup of coffee on an average day (...) Coffee drinkers tend to be older, with 74% of adults aged 55 and older consuming it daily, versus 50% of those aged 18 to 34."[9] *Exercises 5.16 to 5.17 are based on this report.*

5.16 The cited value of 64% is part of

a. the marginal distribution of coffee consumption.
b. the conditional distribution of coffee consumption, given age group.
c. the conditional distribution of age group, given coffee consumption.

5.17 The cited value of 74% is part of

a. the marginal distribution of age group.
b. the conditional distribution of coffee consumption, given age group.
c. the conditional distribution of age group, given coffee consumption.

5.18 To help consumers make informed decisions about health care, the government releases data about patient outcomes in hospitals. A large regional hospital and a small private hospital both serve your community. The regional hospital receives many patients in critical condition because of its state-of-the-art emergency room and diverse set of medical specialties. The private hospital specializes in scheduled surgeries and admits few patients in critical condition. The counts of patients who survived surgery or did not survive surgery are provided for both hospitals, for all surgeries performed in the previous year. The regional hospital had the higher surgery survival rate (the percent of patients still alive six weeks after surgery) for both patients admitted in critical condition and patients coming in for scheduled surgery. Yet, the private hospital had the higher overall survival rate when considering both types of surgery patients together. This finding is

a. not possible: If the regional hospital had had higher survival rates for each type of patient separately, then it must also have had a higher overall survival rate when both types of patient are combined.

b. an example of Simpson’s paradox: The regional hospital’s survival rate was better with each type of patient but it did worse overall because it took in many more patients in critical condition, who had a lower chance of survival due to their condition.

c. due to comparing two conditional distributions that should not be compared.

CHAPTER 5 EXERCISES

5.19 Depression in America. A 2014 Gallup survey on mental health in the United States interviewed a random sample of 173,655 American adults. Here is a table found in the online report of the survey findings.

	Percent currently being treated for depression
All Americans	11
Baby boomers (1946−1964)	14
Generation X (1965−1975)	11
Traditionalists (1900−1945)	9
Millennials (1980−1996)	7

a. Give the marginal distribution of current depression treatment.

b. Explain why the four values in the bottom part of the table do not add to 100% (up to roundoff error).

5.20 Smartphone behavior. A 2015 Pew Research Center survey on technology use asked a random sample of American adult smartphone users if they ever use their smartphone to avoid others around them. The findings are summarized here, by age group.

Age group	**Number surveyed**	**Percent using their smartphone to avoid others around them**
Young adult (18−29)	202	47
Middle age (30−49)	403	32
Older adult (50+)	426	15

a. What proportion of all adult smartphone users in the survey said they use their smartphone to avoid others around them?

b. What does the conditional distribution of smartphone avoidance strategy, given age group, showcase about this new behavior?

5.21 Botox for excessive sweating. Hyperhidrosis is a stressful medical condition characterized by chronic excessive sweating. Primary hyperhidrosis is inherited and typically starts during adolescence. Botox was approved for treatment of hyperhidrosis in adults in 2004. A clinical trial examined the effectiveness of Botox for excessive armpit sweating in teenagers aged 12 to 17. Participants received either a Botox injection or a placebo injection (a simple saline solution) and were examined four weeks later to see if their sweating had been reduced by 50% or more. Here are the study's findings:[10]

	At least 50% reduction		
Treatment	**Yes**	**No**	**Total**
Botox	84	20	104
Placebo	44	64	108
Total	128	84	212

a. Find the marginal distribution of outcome (sweating reduced or not). What does this distribution tell us?

b. How do the outcomes for the teenagers in the study differ depending on which treatment they received? Use conditional distributions as a basis for your answer. What can you conclude about the effectiveness of Botox in treating excessive armpit sweating?

5.22 Weight lifting injuries. Resistance training is a popular form of conditioning aimed at enhancing sports performance. It is widely used by high school, college, and professional athletes, although its use for younger athletes remains controversial. A random sample of 4111 patients between the ages of 8 and 30 admitted to U.S. emergency rooms with the injury code "weightlifting" was obtained.

These injuries were classified as "accidental" if caused by a dropped weight or improper equipment use. The patients were also classified into four age categories. Here is a two-way table of the results:[11]

	Accidental	**Not accidental**
8–13 years	295	102

14–18 years	655	916
19–22 years	239	533
23–30 years	363	1008

Compare the distributions of ages for accidental and nonaccidental injuries. Use percents and draw a bar graph. What do you conclude?

5.23 Fast food for baby. The Gerber company sponsored a large survey of the eating habits of American infants and toddlers. Among the many questions parents were asked was whether their child had eaten fried potatoes on one given day. Here are the data broken down by the children's age range:[12]

	Ate fried potatoes	**Did not**
9–11 months	61	618
15–18 months	62	246
19–24 months	82	234

a. Calculate the conditional distribution of children who ate fried potatoes for each age range. Briefly describe your findings. (The study also found that fried potatoes were actually the most common cooked vegetable in the diet of the 19- to 24-month-old toddlers, way ahead of green beans and peas!)

b. Do you think that the association between age and diet found by this study is evidence that age actually *causes* a change in diet?

5.24 Binge drinking. Binge drinking is the excessive consumption of alcohol in one setting (five or more alcoholic drinks for men; four or more drinks for women due to their typically smaller stature). It is the most common pattern of excessive alcohol use in the United States, according to the CDC. More than 100,000 adults in the National Survey on Drug Use and Health responded to an extensive questionnaire about their alcohol drinking habits in the past year. From the answers, researchers determined who could be medically defined as alcohol-dependent.[13]

a. Figure 5.4 is a graph reproduced from the published survey report. What type of graph is it?

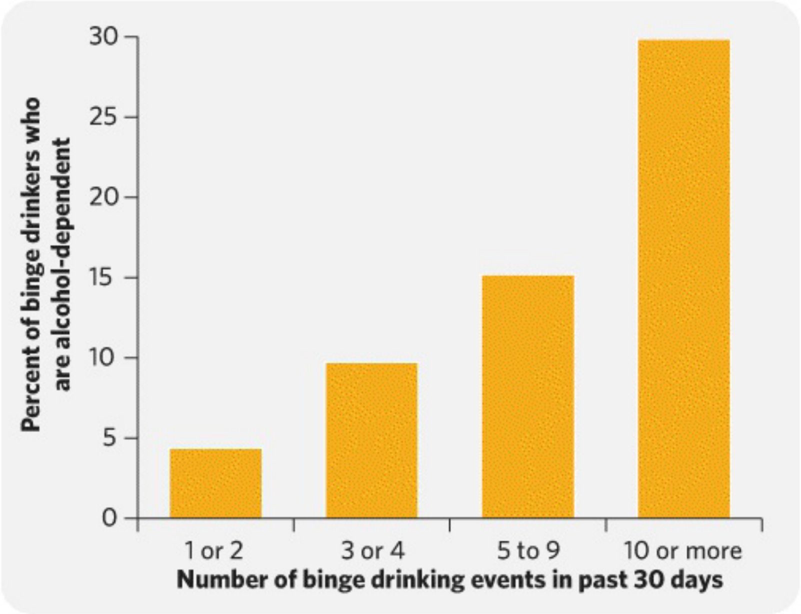


Figure 5.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 5.4 Prevalence of alcohol dependence among U.S. adult binge drinkers, by binge drinking frequency during the past 30 days.

b. Does this graph display a marginal or a conditional distribution? Identify specifically which distribution is displayed (be sure to use actual variable names).

5.25 Preventing neural-tube defects. A key study in the 1990s examined the effectiveness of prenatal vitamin and mineral supplements in preventing neural-tube defects and congenital malformations in newborns. A total of 4156 women planning a pregnancy were given either a combined vitamin/mineral supplement (including folic acid) or a mineral-only supplement to take daily for at least one month before conception. Here is a two-way table of the number of births with a congenital malformation, a neural-tube defect, or neither in both groups:[14]

	Vitamins and minerals	Minerals only

Congenital malformation	28	47
Neural-tube defect	0	6
Neither	2076	1999
Total	2104	2052

Find the conditional distribution of birth outcomes, given the type of prenatal supplement taken. Use your results to describe the effect of adding vitamins to prenatal mineral supplements in terms of newborn health.

5.26 HIV testing in the United States. The CDC recommends that HIV testing be a part of routine clinical care for all sexually active patients. But do people get tested for HIV regularly? The 2012 National Health Interview Survey gives the number of adults in the United States (in thousands) classified by age group and whether they had been tested for HIV in the past 12 months. Here are the data:

	Tested	**Not tested**
15–19 years	2,331	18,964
20–24 years	4,866	15,893
25–29 years	5,023	16,270
30–34 years	3,377	15,039
35–39 years	3,249	17,694
40–44 years	2,425	18,753

Find the conditional distribution of HIV-testing status for each age group, and describe the state of HIV testing in the United States.

5.27 Psoriasis treatment. Psoriasis is an autoimmune disease that manifests as chronic skin inflammation. Researchers recruited adult patients with moderate to severe plaque psoriasis and randomly assigned them to one of four treatment options. Patients were evaluated after twelve weeks of treatment. Here are the findings:[15]

	No visible symptoms	**Symptoms still visible**	**Total**

Placebo (fake treatment)	2	307	309
Ustekinumab	65	235	300
Brodalumab, low dose	157	453	610
Brodalumab, high dose	272	340	612

a. Obtain the conditional distribution of successful outcomes (no more visible symptoms), given the treatment administered.

b. A placebo is a fake treatment designed to resemble an actual treatment (all four treatments here were administered via subcutaneous injections). Did any of the active treatments perform better than the placebo? Why do you think it is important to compare the results from the active treatments with the results when a placebo is administered?

c. Compare the percent of patients showing no more symptoms for each of the three active treatment groups. Was one active treatment more effective than the others in this study?

5.28 Myopia in young children. Inadequate exposure to outdoor lighting conditions may increase the risk of myopia in children. A study compared rates of myopia among randomly selected young children (6 and 7 years old) of Chinese ethnicity living in either Australia or Singapore, because children in Singapore tend to spend substantially less time outdoors than children in Australia. The researchers found that 4 of the 124 children in the Australian sample had myopia, compared with 183 of the 628 children from Singapore.[16]

a. Who are the individuals studied and what information do we have about them? Organize the reported findings in a two-way table of counts.

b. What percent of the children in the study have myopia? What percent of the children from each country have myopia?

c. Would it be appropriate to conclude from this study that spending less time outdoors causes young children of Chinese ethnicity to develop myopia? Explain your reasoning.

d. The researchers enrolled children in both countries who were all 6 to 7 years old and of Chinese ethnicity. What lurking variables were the researchers aiming to avoid with this selection choice? Explain briefly why this choice helps us interpret their findings.

5.29 Type 1 and type 2 diabetes. Type 1 diabetes is usually first diagnosed in children or young adults; it arises when the immune system interferes with the ability of the pancreas to make insulin. Type 2 diabetes can develop at any age, though it is especially common among older individuals; it is an induced insulin resistance, typically associated with excess weight and inactivity. Type 2 diabetes is by far the more common form of diabetes. Records from a diabetes clinic in the United Kingdom show the status (still alive or dead) of patients with long-term diabetes as a function of their diabetes type:[17]

Patients 40 or younger

	Type 1	Type 2
Alive	129	15
Dead	1	0
Total	130	15

Patients older than 40

	Type 1	Type 2
Alive	124	311
Dead	104	218
Total	228	529

a. Compare the survival rates (percents still alive) for type 1 and for type 2 diabetes among the younger patients. Do the same for the older patients. What have you learned from these percents?
b. Combine the data into a single two-way table of patient status (alive or dead) by diabetes type (1 or 2). Now calculate the survival rates for type 1 and for type 2 diabetes among all patients together. Which type of diabetes is associated with the higher survival rate, overall?
c. This study provides an example of Simpson’s paradox. What is the lurking variable here? Explain in simple language how the paradox can happen.

5.30 Obesity and health. Recent studies have shown that earlier reports underestimated the health risks associated with being overweight. The source of this error was overlooking lurking variables. In particular, smoking tends both to reduce weight and to lead to earlier death. Illustrate Simpson’s paradox by giving a simplified version of this situation. That is, create separate two-way tables of overweight (yes or no) by early death (yes or no) for smokers and nonsmokers such that

- overweight smokers and overweight nonsmokers both tend to die earlier than persons who are not overweight,
- but when smokers and nonsmokers are combined into a two-way table of overweight by early death, persons who are not overweight tend to die earlier.

5.31 Python eggs. How is the hatching of water python eggs influenced by the temperature of the snake's nest? Researchers assigned newly laid eggs to one of three temperatures: hot, neutral, or cold. The hot condition duplicates the warmth provided by the mother python. The neutral and cold conditions mimic the mother's absence. Here are the data on the number of eggs and the number that hatched:[18]

	Cold	**Neutral**	**Hot**
Number of eggs	27	56	104
Number hatched	16	38	75

a. Notice that this is not a two-way table! Explain why and how you could use the data to create a two-way table.
b. The researchers anticipated that eggs would hatch less well at cooler temperatures. Do the data support that anticipation? Follow the four-step process (page 58) in your answer.

5.32 Do angry people have more heart disease? People who get angry easily tend to have more heart disease—at least, that was the conclusion of a study that followed a random sample of 12,986 people for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data compiled at the end of the study for the 8474 people in the sample who had normal blood pressure.[19] CHD stands for "coronary heart disease," which includes people who had heart attacks and those who needed medical treatment for heart disease.

	Low anger	**Moderate anger**	**High anger**	**Total**
CHD	53	110	27	190

No CHD	3057	4621	606	8284
Total	3110	4731	633	8474

Do these data support the study's conclusion about the relationship between anger and heart disease? Follow the four-step process as in Example 5.3.

5.33 Characteristics of a forest. Forests are complex, evolving ecosystems. For instance, pioneer tree species can be displaced by successional species better adapted to the changing environment. Ecologists mapped a large Canadian forest plot dominated by Douglas fir with an understory of western hemlock and western red cedar. The following two-way table records all the trees in the plot by species and by life stage. The distinction between live and sapling trees corresponds to live trees that were taller or shorter than 1.3 meters, respectively.[20]

	Dead	**Live**	**Sapling**	**Total**
Western red cedar	48	214	154	416
Douglas fir	326	324	2	652
Western hemlock	474	420	88	982
Total	848	958	244	2050

What can you tell from these data about the current composition and the evolution of this forest? Which tree species appears to be taking over and becoming dominant? Follow the four-step process as in Example 5.3.

5.34 Complications of bariatric surgery. Bariatric surgery, or weight-loss surgery, includes a variety of procedures performed on people who are obese. Weight loss is achieved by reducing the size of the stomach with an implanted medical device (gastric banding), by removing a portion of the stomach (sleeve gastrectomy), or by resecting and rerouting the small intestines to a small stomach pouch (gastric bypass surgery). Complications can occur with any of these methods. Serious complications include life-threatening, permanently disabling, and fatal outcomes. Here is a two-way table for data collected in Michigan over several years giving counts of non-life-threatening complications, serious complications, and no complications for these three types of surgeries:[21]

	Complications			
	Non-life threatening	Serious	None	Total
Gastric banding	81	46	5253	5380
Sleeve gastrectomy	31	19	804	854
Gastric bypass	606	325	8110	9041

What do the data say about differences in complications for the three types of surgeries? Follow the four-step process.

5.35 Smoking cessation. A large randomized trial was conducted to assess the efficacy of Chantix for smoking cessation compared with bupropion (more commonly known as Wellbutrin or Zyban) and a placebo. Chantix is different from most other quit-smoking products in that it targets nicotine receptors in the brain, attaches to them, and blocks nicotine from reaching them. In contrast, bupropion is an antidepressant often used to help people stop smoking. In the clinical trial, generally healthy smokers who smoked at least 10 cigarettes per day were assigned at random to take Chantix (n=352),$(n = 352)$, bupropion (n=329),$(n = 329)$, or a placebo (n=344).$(n = 344)$. The response variable was the continuing cessation from smoking for weeks 9 through 12 of the study. Here is a two-way table of the results:[22]

	Treatment		
	Chantix	Bupropion	Placebo
No smoking in weeks 9–12	155	97	61
Smoked in weeks 9–12	197	232	283

How does whether a subject smoked in weeks 9 to 12 depend on the treatment received? Follow the four-step process (page 58).

CHAPTER 6 Samples and Observational Studies

EBFoto/Depositphotos

IN THIS CHAPTER WE COVER...

- Observation versus experiment
- Sampling
- Sampling designs
- Sample surveys
- Cohorts and case-control studies

Statistics, the science of data, provides ideas and tools that we can use in many settings. Raw data should always be carefully examined for trends and deviations from these trends using the tools of exploratory data analysis. But how do we *collect data* that can help us answer interesting scientific questions?

Suppose our question is “What percent of adult Americans use complementary and alternative medicine (CAM), such as herbal products and meditation used for health purposes?” To answer the question, we ask individuals who are age 20 or older whether they have used any form of CAM in the past 12 months. We can’t realistically ask every adult in America, so we put the question to a *sample* chosen to represent the entire adult *population.* How shall we choose such a sample? In this chapter we examine how to choose appropriate samples for observational studies.

OBSERVATION VERSUS EXPERIMENT

We collect data to understand phenomena or answer scientific questions. There are two approaches for collecting data. One is the *observational study*, in which the subject of our scientific inquiry is disturbed as little as possible by the act of gathering information. Observational studies can answer questions such as "What is the range of gestation times in the prairie dog?" and "What is the proportion of Americans who are overweight?" The other approach is that of the *experiment*. In doing an experiment, we actively impose some *treatment* or condition in order to observe the *response*. Experiments can answer questions such as "Does aspirin reduce the chance of a heart attack?" and "Does ginkgo biloba enhance memory?" Experiments and observational studies provide useful data only when properly designed. The distinction between experiments and observational studies is one of the most important ideas in statistics.

OBSERVATION VERSUS EXPERIMENT

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses. The purpose of an observational study is to describe and compare existing groups or situations.

An **experiment** deliberately imposes some treatment on individuals in order to record their responses. The purpose of an experiment is to study whether the treatment causes a change in the response.

Observational studies are essential sources of data about topics ranging from the heights of adult Americans to the behavior of animals in the wild. But an observational study, even one based on a statistical sample, is a poor way to gauge the effect of an intervention. To see the response to a change, we must actually impose the change. When our goal is to understand cause and effect, experiments are the only source of fully convincing data.

EXAMPLE 6.1 The ups and downs of hormone replacement

Should women take hormones such as estrogen after menopause, when natural production of these hormones ends? In 1992, several major medical organizations said “yes.” In particular, women who took hormones seemed to reduce their risk of a heart attack by 35% to 50%. The risks of taking hormones appeared small compared with the benefits.

The evidence in favor of hormone replacement came from a number of observational studies that compared women who were taking hormones on their own accord with others who were not. But women who elect to take hormones are most likely quite different from women who do not: They may be better informed and see their doctors more often, and they may do many other things to maintain their health. So, is it surprising that they have fewer heart attacks?

Experiments don’t let participants decide what to do; instead they assign them to specific conditions. The Women’s Health Initiative (WHI) trial sponsored by the National Institutes of Health assigned women either to hormone replacement or to dummy pills that look and taste the same as the hormone pills. The assignment was done randomly, so that all kinds of women were equally likely to get either treatment. In 2002, the WHI trial published its first results, which indicated that women who took hormones had a *higher* incidence of cardiovascular disease and breast cancer. Taking hormones after menopause quickly fell out of favor. This first WHI study, however, had focused on older women, with an average age of 63 years. When a follow-up WHI trial of women in their 50s was published in 2007, it showed that younger women taking hormone therapy had lower levels of calcium deposits in their arteries, which may lower their risk of heart disease.[1]

When we simply observe individuals, there are many more variables than the ones we record, and some of those may affect our response variable. We call them **lurking variables.** Lurking variables greatly complicate our ability to interpret findings because their effect and the effect of the explanatory variable are *confounded* (mixed up).

CONFOUNDING

Two variables (either explanatory variables or lurking variables) are **confounded** when their effects on a response variable cannot be distinguished from each other.

Observational studies of the effect of one variable on another often fail to demonstrate causality because the explanatory variable is confounded with lurking variables. We will see in the next chapter that well-designed experiments take steps to defeat confounding, although they are not entirely immune to lurking variables. We saw in Example 6.1 that the WHI trial initially failed to account for the women's age.

EXAMPLE 6.2 Wine, beer, or spirits?

In Example 4.7 (page 108) we saw that the observed association between moderate use of alcohol and better health exists in part because some non-drinkers choose to avoid alcohol due to preexisting health conditions.

Among people who do drink alcohol in moderation, observational studies found that drinking wine rather than beer or spirits is associated with better health. But people who prefer wine are different from those who drink mainly beer or stronger spirits. Moderate wine drinkers as a group are richer and better educated. They eat healthier food and are less likely to smoke. The explanatory variable (Which type of alcoholic beverage do you drink most often?) is confounded with many lurking variables (education, wealth, diet, and so on). A large study therefore concluded: "The apparent health benefits of wine compared with other alcoholic beverages, as described by others, may be a result of confounding by dietary habits and other lifestyle factors."[2] Figure 6.1 shows the confounding in visual form.

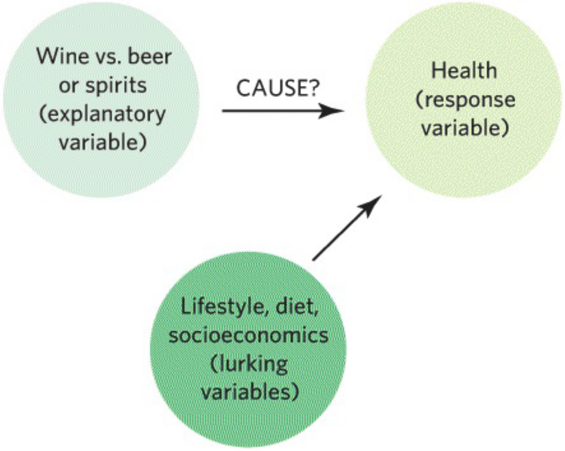

Wine vs. beer or spirits (explanatory variable)
CAUSE?
Health (response variable)
Lifestyle, diet, socioeconomics (lurking variables)

Figure 6.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, ©
2018 W. H. Freeman and Company

FIGURE 6.1 Confounding: We cannot distinguish the effects of what people drink from the effects of their overall lifestyle, diet, and socioeconomics.

APPLY YOUR KNOWLEDGE

6.1 Brain asymmetry. A study aims to determine the percent of adults whose right cerebral hemisphere is larger than their left hemisphere. A group of individuals age 20 and older representative of the American adult population is asked to come to a brain-imaging facility to have their brains scanned. Each individual’s scan is then analyzed to identify whether the right or left cerebral hemisphere is larger. Is this an observational study or an experiment? Explain your answer.

6.2 Physician-assisted suicide. The Gallup polling agency regularly investigates Americans’ attitude toward the issue of physician-assisted suicide. To investigate how phrasing affects opinions, Gallup creates two versions of the question. Version A asks, “When a person has a disease that cannot be cured, do you think doctors should be allowed by law to end the patient’s life by some painless means if the patient and his or her family request it?” Version B instead asks, “When a person has a disease that cannot be cured and is living in severe pain, do you think doctors should or should not be allowed by law to assist the patient to commit suicide if the patient requests it?” Using random assignment, roughly half the study participants are asked version A of the question and the other half are asked version B.[3] Is this an observational study or an experiment? Explain your answer.

6.3 Hydration and obesity. A study examined a large nationally representative sample of American adults age 18 to 64 years. The primary outcome of interest was body mass index (BMI), which was also used to categorize participants as obese ($\text{BMI} \geq 30$$\text{BMI} \geq 30$) or not obese ($\text{BMI} < 30$$\text{BMI} < 30$). In the study, inadequate hydration (as evidenced by a urine osmolality ≥ 800 mOsm/kg as evidenced by a urine osmolality $\geq 800\,\text{mOsm/kg}$) was associated with higher BMI and obesity. The published findings include this statement: “Although inadequate intake of water among obese adults may explain the observed findings, differential consumption of food with high water content may also contribute to the relationship between inadequate hydration and elevated BMI.”[4]

a. Is this study observational or experimental?
b. What variables recorded for each participant are described here?
c. What lurking variable cited here is confounded with the explanatory variable?
d. Would it be appropriate, based on this study alone, to advise obese patients that drinking more water will help them lower their BMI? Explain your answer.

SAMPLING

The Centers for Disease Control and Prevention (CDC) wants to know the DNA sequences of the flu viruses circulating around the world to help prepare next year's flu vaccine. The Census Bureau wants to know what percent of children age 16 and younger have received all recommended immunizations. In both cases, we want to gather information about a very large group of individuals. Time, cost, and practicality forbid collecting data about each flu case in the world or each American child's immunization records. Instead, we gather information about only part of the group, a *sample*, so that we can then draw conclusions about the whole, the *population* of interest.

POPULATION, SAMPLE, SAMPLING DESIGN

The **population** in a statistical study is the entire group of individuals (not necessarily people) about which we want information.

A **sample** is the part of the population from which we actually collect information. We use a sample to draw conclusions about the entire population.

A **sampling design** describes exactly how a sample is chosen from the population.

We often draw conclusions about a whole on the basis of a sample. Everyone has sipped a spoonful of soup and judged the entire bowl on the basis of that taste. When your doctor is concerned about your white blood cell count, you give a “blood sample” to see how low your white cell count really is. But the bowl of soup and your blood are quite homogeneous, so that the taste of a single spoonful represents the whole soup and a small vial of your blood represents your whole blood composition at the time of extraction. These are relatively easy sampling designs.

Conclusions reached from analyzing sample data extend only to a population of individuals similar to those sampled, so it is important that the sample truly represent the intended population. Choosing a representative sample from a large and varied population is not so simple. The first step in a proper sample design is to say exactly *what population* we want to describe. The second step is to say exactly *what we want to record*—that is, to give exact definitions of our variables.

EXAMPLE 6.3 The Current Population Survey

The Current Population Survey (CPS) is a highly important sample survey conducted by the U.S. government. The CPS contacts approximately 60,000 households each month. The results are used to determine the monthly unemployment rate and various demographic characteristics. Supplemental questionnaires on a variety of topics are added at regular intervals. They are designed to update other government surveys such as the National Health Interview Survey, which is conducted on a yearly basis.

One question of interest to the government is the proportion of current U.S. smokers who attempted to quit smoking within the previous 12 months. The first step in producing data for this question is to specify the population we want to describe. Which age groups will we include? Will we include undocumented immigrants or people in prisons? The CPS defines its population as all U.S. residents (whether citizens or not) who are at least 15 years of age, who are civilians, and who are not in an institution such as a prison or a nursing home. Persons from all 50 states are considered, representing both genders and all ethnic groups.

The second step involves choosing what precisely we want to measure. What does it mean to be a "current smoker"? Should we include individuals who smoke only occasionally? What constitutes an "attempt to quit smoking"? The CPS defines current smokers as individuals who currently smoke either every day or some days. If you are a current smoker, the interviewer then goes on to ask about quitting attempts. A quitting attempt is defined as an attempt to quit smoking during the past year that lasted for 24 hours or more.

EXAMPLE 6.4 Animal studies

Tree shrews, *Tupaia belangeri*, are small omnivorous mammals that are phylogenetically related to primates. Do tree shrews exhibit a paw preference when grasping food?[5] The first step in producing data to answer this question is to specify the population we want to describe. This would be the population of all tree shrews—a population naturally found in Asian tropical forests. What are researchers' options for obtaining a sample of live animals? They can capture tree shrews in the wild or obtain tree shrews from a dedicated breeding facility. The researchers in this study chose to obtain 36 tree shrews from a breeding facility run by the University of Veterinary Medicine, Hanover, Germany. The advantage is that their population of interest is similar to that studied by other researchers in their field. The disadvantage is that these animals are not representative of the population of wild tree shrews. The second step involves choosing what we want to measure: How do we assess "paw preference"? The researchers observed the tree shrews during food-grasping tasks and recorded which paw was used for grasping. They used this information to compute a quantitative "pawedness index" reflecting the relative use of the right and left paws during numerous grasping attempts; they then used this index value to label each animal as having a right, left, or ambidextrous paw preference.

EBFoto/Depositphotos

Another study examined the biomechanics of jumping in the hedgehog flea, *Archaeopsyllus erinacei*.[6] The target population is all adult hedgehog fleas (only the adults jump), but how do you sample from this population? The researchers obtained 10 adult specimens from hedgehogs at St. Tiggywinkles Wildlife Hospital Trust in the United Kingdom. These fleas are representative of wild fleas in the area around the wildlife hospital, but there is no guarantee that they are truly representative of all hedgehog fleas everywhere. One aspect of jump examined in this study was takeoff angle. The researchers filmed the fleas in slow motion during spontaneous jumps and recorded the takeoff angle of each jump.

These two examples bring up an important issue. *Not all statistical studies use samples that are drawn directly from the entire population of interest.* Studies of animal behavior, for example, cannot realistically capture animals from their entire wildlife habitat. Whether the results may also apply to a larger population (the entire tree shrew breeding colony in Hanover or even the entire population of tree shrews) is a biological argument, not a statistical one. We will discuss this issue further in the next chapter.

CAUTION

APPLY YOUR KNOWLEDGE

6.4 Sampling members. A health care provider wants to rate its members' satisfaction with physical therapy. A questionnaire is mailed to 800 members of the health plan selected at random from the list of members who were prescribed physical therapy in the past 12 months. Only 212 questionnaires are returned.

a. What is the *intended* population in this study? Be careful: What group does the health care provider *want information about?*

b. What is the sample? Be careful: From what group does the provider *actually obtain information?*

6.5 Manatee deaths. Manatees are large aquatic mammals found primarily in the rivers and estuaries of Florida. Because they are an endangered species, the Florida Fish and Wildlife Conservation Commission tracks all manatee deaths in the state and establishes a probable cause of death for each carcass. In 2016, the commission recorded 520 manatee deaths.[7] Does this number represent a sample of manatee deaths or does it represent the actual manatee population in Florida?

SAMPLING DESIGNS

The purpose of a sample is to give us information about a larger population. The process of drawing conclusions about a population on the basis of sample data is called statistical **inference** because we *infer* information about the population from what we *know* about the sample.

Some important attributes of the sample can make the process of inference either easier or more challenging. Example 6.5 illustrates these possibilities using an archery analogy.

EXAMPLE 6.5 Analogy for sample attributes

Imagine an archery range where people come to practice. At the end of a training session, the iconic multicolor archery target is removed but perhaps you can still see the holes left by the arrows. Think of these holes as actual data points in a sample. Figure 6.2 shows four scenarios.

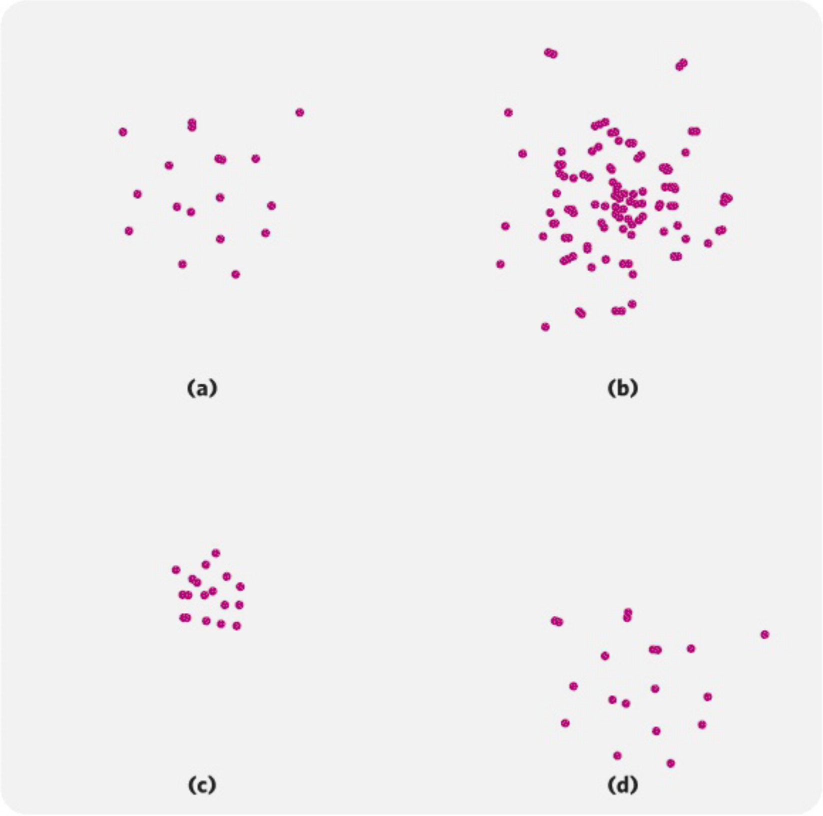


Figure 6.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 6.2 Analogy for taking samples. Four samples are shown. Imagine guessing the location of the center of the population from the center of the sample distribution.

Our goal is to guess, with a given level of certainty, the location of the target's center where the archer was aiming. In our archery analogy, this would be the center of the population. Draw a circle in each scenario where you are fairly certain that the center of the target should be, based on the pattern of points. Then look at Figure 6.3 to see whether your guess was correct. Comparing these scenarios highlights important attributes of samples:

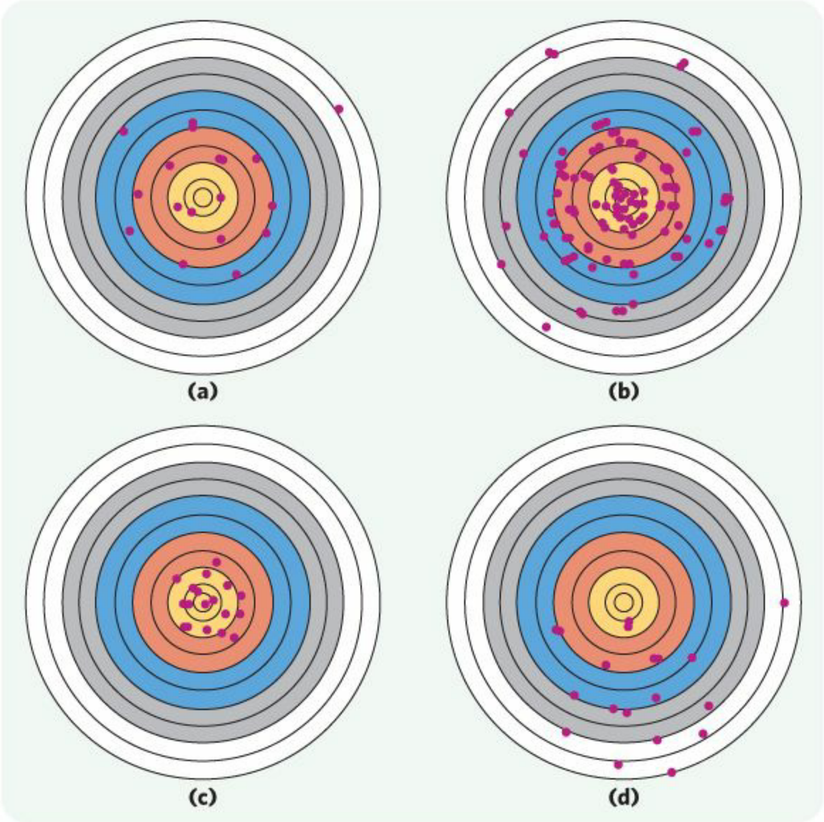


Figure 6.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 6.3 The samples from Figure 6.2 are now superimposed on archery targets representing the target population. Sample (d) in this analogy is a biased sample.

- (a) versus (b): More arrows were shot in scenario (b), so there are a lot more data points. More information makes it easier to pinpoint the likely center of the target.

- (a) versus (c): The same number of arrows were shot in both scenarios, but scenario (c) shows less variability (perhaps a more practiced archer). Less variability makes it easier to pinpoint the likely center of the target.
- (a) versus (d): Both patterns of points are identical. Yet, the systematic downward misses (perhaps from an improperly adjusted bow) in scenario (d) make it practically impossible to correctly guess the center of the target.

This analogy illustrates visually the impact of important sample attributes on inference. We can accurately perform inference for any reasonable sample size and data variability, but larger samples and lower variability lead to more precise inference. In contrast, systematic errors in the data collection process (*bias*) are likely to lead to a wrong conclusion. A large part of statistics consists of identifying possible sources of bias and designing data collection to avoid them.

BIAS

The design of a statistical study is **biased** if it systematically favors certain outcomes.

Because of potential bias, not all sampling designs are appropriate for statistical purposes.

Poor sampling designs The easiest—but not the best—sampling design just chooses individuals close at hand. This is called a **convenience sample,** and it often produces unrepresentative data. If we are interested in finding out what percent of adults take vitamin and mineral supplements, for example, we might go to a shopping mall and ask people we meet there. However, a shopping mall sample will almost surely overrepresent middle-class and retired people and underrepresent the poor. This pattern will recur almost every time we take such a sample. That is, the source of bias is a systematic error caused by a bad sampling design, not just bad luck on one sample. The outcomes of shopping mall surveys will repeatedly miss the truth about the population in the same ways.

Another biased sampling design, the **voluntary response sample** (also called volunteer or self-selected sample), lets individuals choose whether to participate in the study. Opt-in polls (write-in, call-in, or online quick votes) are examples of voluntary response samples. They are not scientific polls. Instead, these types of samples are biased because people with strong opinions are most likely to respond. The problem is that *people who take the trouble to respond to an open invitation are usually not representative of any clearly defined population.* Voluntary response samples may also allow individuals to submit any number of entries, further exacerbating the potential for bias.

CAUTION

Unfortunately, convenience and voluntary response samples have become very popular with the media because they are cheap and often produce (misleadingly) sensational results.[8]

EXAMPLE 6.6 Opt-in poll: “Sex and Love”

A study analyzed the answers of 60,058 heterosexual participants to an open poll titled “Sex and Love” posted for 10 days on the official website of NBC News.[9] Because anyone could participate, we have no way of knowing what types of individuals chose to participate or even if they participated only once. There is therefore no larger, clearly defined population here about which to infer.

The researchers acknowledged that their sample “was not nationally representative” and described two main reasons why. They cited research showing that voluntary response samples recruited online “tend to include participants who are relatively more educated and have higher income than the national population.” They also pointed out the possibility “that the survey title *Sex and Love* appealed to people with more liberal attitudes toward sex.”

In a voluntary response sample, people choose whether to respond to an open call for participation. In a convenience sample, the interviewer makes the choice out of convenience. In both cases, personal choice introduces bias. The statistician’s remedy is to allow impersonal chance to select the sample.

Simple random samples A sample chosen by chance allows for neither favoritism by the sampler nor self-selection by respondents. Choosing a sample by chance mitigates bias by giving all individuals an equal chance to be chosen. Rich and poor, young and old, male and female—all have the same chance to be in the sample. A sample chosen by chance is called a probability sample. The most common **probability sampling** designs are simple random sampling, stratified random sampling, and multistage random sampling.

SIMPLE RANDOM SAMPLE

A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be the sample actually selected.

An SRS not only gives each individual an equal chance to be selected, but also gives every possible sample of size n an equal chance to be selected. When you think of an SRS, picture drawing names from a hat to remind yourself that an SRS doesn’t favor any part of the population. That’s why an SRS is a better method of selecting samples than convenience or voluntary response sampling.

Of course, writing names on slips of paper and drawing them from a hat is slow and inconvenient, especially for large populations (for example, all American adults). In practice, most people use technology. The *Simple Random Sample* applet makes choosing an SRS very fast. If you don’t use the applet or other software, you can use a **table of random digits,** like Table A at the back of the book, which consists of a long random sequence of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Exercise 6.7 guides you through the process.

APPLET

We can trust that an SRS avoids selection bias, because it uses impersonal chance to select the individuals in the sample. Voluntary opt-in polls, like the one in Example 6.5, and mall interviews also produce samples. But we can't trust results from these samples, because they are chosen in ways that invite bias.

Obtaining a true SRS can be challenging because it requires that we be able to identify and label all the individuals in the population. Variations on the approach may also provide samples selected at random to mitigate bias.

EXAMPLE 6.7 Examples of random selection

A health care organization has comprehensive records for all 4000 of its patients, including their billing addresses. To find out how satisfied its patients are with the care provided, a manager assigns the patients in the company database a number from 1 to 4000 and uses software to select 100 patients at random who will receive a customer satisfaction survey in the mail. The result is an SRS of 100 patients.

Researchers in Guam examined 504 brown treesnakes, an invasive species causing substantial environmental damage to the island. How were these 504 snakes "selected"? The researchers placed, at regular intervals in a large forested area, standard brown-treesnake traps with a mouse bait.[10] The result is a reasonable approximation of an SRS because every brown treesnake in the forested area should have the same chance of ending in the sample.

Researchers studied the contents of 800 vaccine-related posts on the social media site Pinterest. How were these posts selected? On three days in March 2014, the researchers chose every fifth post containing one of these keywords: vaccination, vaccine, vaccines, and vaccinate.[11] This is a **systematic** random sample, a form of probability sampling in which every kth element is selected in sequence after randomly choosing a starting point. The population here is all Pinterest posts on these days with one of these four keywords. As long as there is no consistent pattern in the order of the posts, this approach offers a random sample without built-in bias in the same way that an SRS would.

Other probability sampling designs The use of chance to select a sample is the essential principle of statistical sampling. Some probability sampling designs (such as an SRS) give each member of the population an equal chance to be selected. This may not be true in more elaborate sampling designs.

For example, it is common to sample distinct groups within the population separately, and then to combine these samples. This is the idea of a **stratified** random sample. It can be useful when the population contains groups that are more rare or harder to reach than others. For example, a study of exercise habits in American adults may use a stratified random sample, selecting individuals of different ethnic groups separately, to ensure inclusion of majority and minority groups according to their proportions in the population. Inversely, a study of health care use in different ethnic groups may use a stratified random sample to ensure the selection of a large enough number of individuals from minority groups so that meaningful comparisons can be made between the groups. In either case, a stratified random sample provides one additional level of control over the selection of individuals in the sample, compared with a straightforward SRS.

Nationwide or statewide governmental surveys often use **multistage** samples for reasons of practicality. These samples typically involve choosing SRSs within SRSs. For example, the National Youth Tobacco Survey selects first an SRS of counties nationwide; then for each chosen county, an SRS of schools is chosen; lastly, within each chosen school, an SRS of classrooms is drawn. This is a more practical and cost-effective method than directly selecting an SRS of students from all over the United States.

Many approaches have been developed to select useful and reliable probability samples, some of them quite complex. Large-scale sample surveys, for example, typically use stratified, multistage samples that combine SRSs from each stage of the sampling process. Analysis of data from sampling designs more complex than an SRS takes us beyond basic statistics and will not be treated further in this book. Nevertheless, the SRS remains the building block of more elaborate designs, and analysis of other probability designs differs more in complexity of detail than in fundamental concepts.

APPLY YOUR KNOWLEDGE

6.6 Sampling on campus. Your college wants to know students' perceptions of the nutritional quality of campus food. It isn't practical to contact all students.

a. Give an example of a way to choose a sample of students that is poor practice because it depends on voluntary response. What type of bias could you expect?
b. Give an example of a bad way to choose a sample of students that doesn't use voluntary response. What type of bias could you expect?
c. Give an example of a good way to choose a sample of students that would not introduce bias.

6.7 Sampling the forest: a simple random sample. To gather data on a 1200-acre pine forest in Louisiana, the U.S. Forest Service laid a grid of 1410 equally spaced circular plots over a map of the forest. A ground survey visited a sample of 10% of these plots.[12] Selecting an SRS is a two-step process. **Step 1. Label:** Give each member of the population a unique numerical label of fixed length. **Step 2. Select:** Use software or a table of random digits to pick nn of the individuals in the labeled population.

a. *Using a table of random digits.* First, label the plots using numerical labels starting with the number 1. Make sure that all plot labels have the same length (think about how many digits are needed for the largest numerical label, in this case 1410). Second, go to line 105 in Table A at the back of the book. Ignore the spacing between digits (their only purpose is readability), and read the digits in groups of the same length as your numerical labels. The first three groups that match a plot label make up the beginning of your sample. Ignore any repeated labels, because you don't want to choose the same plot twice.
b. Use the software of your choice to select the first 3 plots in an SRS of 141 plots. For example, the function "= RANDBETWEEN(bottom, top)" in Excel returns an integer

between a bottom and a top number you select.

6.8 America’s State of Mind Report: a stratified random sample. Medco Health Solutions is a prescription drug insurance provider with approximately 65 million members. Medco took a random sample of its members to analyze the utilization of mental health–related medications among the insured population. Sampling was stratified by age, sex, and region to match the demographics of Medco’s members. Explain how a stratified sample is different from a simple random sample and what both types of samples have in common.

6.9 The National Health Interview Survey: a multistage random sample. The CDC conducts the yearly National Health Interview Survey (NHIS). Here is a description of its methodology: “To achieve sampling efficiency and to keep survey operations manageable, cost-effective, and timely, the NHIS survey planners used multistage sampling techniques to select the sample of persons and households for the NHIS.” Explain how a multistage sample is different from a simple random sample and what both types of samples have in common.

SAMPLE SURVEYS

A sample survey is an observational study that relies on a random sample drawn from the entire population of interest at one point in time. Sample surveys most often assess the characteristics or opinions of people, and have a wide array of applications. The Current Population Survey, described in Example 6.3, is a comprehensive survey of households from all parts of the United States (approximately 100 million U.S. households in all). Opinion polls are sample surveys that cover all sorts of topics and typically use voter registries or telephone numbers to select their samples. Observational studies that collect data about a population at one point in time are sometimes called **cross-sectional studies.** In epidemiology, cross-sectional sample surveys are used to estimate the proportion of individuals in the population with various medical conditions and diseases.

Sample surveys typically collect lots of data for each individual sampled. In addition to the main questions of interest, personal information about the respondents is often recorded, such as gender, age, and sociodemographics. This enables the comparison of results between subgroups of individuals at the data analysis stage.

Random selection of the individuals surveyed eliminates bias in the choice of a sample from a list of the population. In practice, this process can be quite challenging. To begin, we need an accurate and complete list of the population (often called a sampling frame). Because such a list is rarely available, most samples suffer from some degree of **undercoverage**. Undercoverage (or *coverage bias*) is a form of selection bias that occurs when some groups in the target population are left out of the process of selecting the sample. If we aim to study the American population, then using a sample survey of households will miss not only homeless people but also prison inmates, individuals in long-term care facilities, and students in dormitories. Conducting the survey by calling landline telephone numbers will miss households that have only cell phones as well as households without a phone. The results of national sample surveys will, in turn, have some bias if the people not covered differ from the rest of the population.

EXAMPLE 6.8 Cell phones and undercoverage

Random digit dialing (RDD) is a common and inexpensive computer-based method of contacting a random sample of people by telephone. However, federal regulations forbid RDD technology for dialing cell phone numbers, which must instead be dialed by hand at greater expense.[13] But can we conduct a reliable telephone survey of the American population using landline numbers only?

This was the issue the CDC considered in 2003 when it planned its first survey of telephone access. The ideal tool for this inquiry was the CDC's own National Health Interview Survey (NHIS), because it interviews participants in person, regardless of their telephone access or lack thereof. The data showed that, in early 2003, 3% of households in the United States had a cell phone but no landline phone ("wireless only" households). This estimate climbed to 25% by early 2010, and to 51% by the end of 2016.[14] Therefore, telephone interviews conducted exclusively via landlines after 2003 would have missed an ever-growing segment of the population.

What would be the impact of excluding wireless-only households from telephone surveys? The NHIS shows that wireless-only individuals differ in a number of ways from the general population. They tend to be younger (in 2016, 73% of adults age 25 to 29 were wireless-only, compared with only 24% of those age 65 and older) and to have a lower socioeconomic status. Their ethnic distribution is different (non-Hispanic whites are less likely to live in wireless-only households). They also differ on several health indicators: Wireless-only individuals are more likely to smoke, to drink excessively, and to be physically active, and they are less likely to have diabetes. Therefore, telephone surveys relying exclusively on landline phones would sample from a population substantially different from the entire U.S. population.

So what is the solution? At first, the CDC simply *adjusted* its landline survey results to compensate for the bias introduced by undercoverage. For example, if a sample contained too few young adults, the responses of the young adults who did respond were given extra weight. When the share of wireless-only households became too substantial to justify this kind of adjustment, the CDC switched to using both wireless and landline numbers in its telephone surveys, in proportions reflecting their respective adoption in the target population (a "dual frame" approach). This approach requires some adjustment to compensate for individuals with both a landline and a cell phone, who would otherwise be twice as likely to be selected (*overcoverage*). Many other polling organizations have since followed the CDC's lead, regularly increasing the proportion of cell phone numbers in their telephone surveys.

Another potential source of bias in most sample surveys is **nonresponse,** which occurs when a selected individual cannot be contacted or refuses to participate. Nonresponse to sample surveys is often very large, even with careful planning and several callbacks. If the people contacted differ from those who are rarely available or who refuse to answer questions, some bias will be introduced.

EXAMPLE 6.9 How bad is nonresponse?

The Census Bureau's American Community Survey (ACS) is a monthly survey mailed each time to a new random sample of nearly 300,000 U.S. addresses (households and group housing). Response to the survey is mandatory. Failure to reply triggers telephone and in-person follow-ups. The ACS has the lowest nonresponse rate of any poll we know, consistently well below 5% historically.[15] Why so much effort? The Census Bureau states that the information collected helps to determine the allocation of more than $400 billion of federal funding each year and offers reliable data for purposes ranging from business decisions to community emergency planning.

The CDC's National Health Interview Survey, described in Example 6.8, interviews participants in person. Its nonresponse rate was 30% in 2015.[16] The University of Chicago's General Social Survey (GSS) is the nation's most important social science survey; it is conducted in person every two years. The GSS contacts its sample in person, and it is run by a university. The 2014 survey had a rate of nonresponse of 31%.[17]

What about opinion polls by news media and opinion-polling firms?

> Few disclose their rates of nonresponse openly. A task force for the American Association for Public Opinion obtained response rate data from eight large and prestigious commercial and nonprofit research firms. Combined, their 2015 rate of nonresponse was 91% for individuals contacted via landline and 93% for those contacted via cell phone.[18]

With large rates of nonresponse comes a potential for bias. Research into survey methodologies attempts to assess the extent of this bias, if any. It is easy enough to compare the basic sociodemographic characteristics of survey respondents with the corresponding characteristics of the target population. When a mismatch occurs, pollsters can *adjust* responses accordingly. It is a much bigger challenge, however, to determine how the opinions of individuals who responded might differ from the opinions of those who did not (since, by definition, those who did not respond did not offer their opinion).[19]

Do not confuse nonresponse in a probability sample with opt-in polls that use a voluntary response sample. Surveys using a probability sample may suffer bias from nonresponse, but this bias can be alleviated to some extent by weighting various groups to make the sample more closely representative of the target population. In contrast, with opt-in polls that rely on voluntary response, we never know who the respondents are and what kind of population they might represent—some individuals may even participate more than once (there are documented examples of thousands of online submissions coming from the same computer[20]). Nothing can be done with voluntary response samples to make the respondents truly representative of the intended target population.

In addition to selection bias and nonresponse bias, the behavior of the respondent or of the interviewer can cause **response bias** in sample results. Drinking a lot of alcohol and having many sexual partners, for example, typically have negative connotations, and many interviewees understate their responses to questions on these topics. A study contrasting self-reports of height and weight with actual measurements showed a tendency to report values that would be considered slightly more desirable.[21]

Responses may also be influenced by whether the interviewer is male or female or from one ethnic group or another. Answers to questions that require recalling past events are often inaccurate because of faulty memory. For example, many people “telescope” events in the past, bringing them forward in memory to more recent time periods. “Have you visited a dentist in the last six months?” will often draw a “yes” from someone who last visited a dentist eight months ago.[22] Careful training of interviewers and careful supervision to avoid variation among the interviewers can reduce response bias. Good interviewing technique is another aspect of a well-done sample survey.

The **wording of questions** can strongly influence the answers given to a sample survey. Confusing or leading questions can introduce strong bias, and even minor changes in wording can change a survey’s outcome. For example, only 13% of Americans surveyed think we are spending too much on “assistance to the poor,” but 44% think we are spending too much on “welfare.”[23] The order in which questions are asked can make a difference, too.

EXAMPLE 6.10 Are you happy?

Ask a sample of college students these two questions:

> “How happy are you with your life in general?” (Answers on a scale of 1 to 5)
>
> “How many dates did you have last month?”

The correlation between answers is $r=-0.012$$r = -0.012$ when asked in this order. It appears that dating has little to do with happiness. Reverse the order of the questions, however, and $r=0.66$.$r = 0.66$. Asking a question that brings dating to mind makes dating success a big factor in happiness.

Don’t trust the results of a sample survey until you have read the exact questions asked. The amount of nonresponse and the date of the survey are also important. A trustworthy survey must go beyond a good sampling design.

CAUTION

APPLY YOUR KNOWLEDGE

6.10 Internet surveys. Given the increasing costs associated with telephone surveys, more pollsters are turning to the internet to reach their survey participants. Survey scientists investigated the ways in which such surveys might differ from telephone surveys. You can read the whole report, written for the general public, at www.pewresearch.org/2015/09/22/coverage-error-in-internetsurveys.

a. The report states that "Pew Research Center telephone surveys have documented the rise in internet adoption, which has grown from 14% of the U.S. adult population in 1996 to 89% today." If a sample survey was conducted exclusively online, from what type of bias would it most likely suffer? Explain your answer.

b. The researchers examined a survey of a large, nationally representative sample in which most individuals answered the questions online but those who were unable or unwilling to take surveys online were interviewed by mail. The survey asked respondents if they have "looked up health information online," to which 63% answered yes. However, within the subset of participants who had answered the survey online, 68% said yes. What might explain the difference between these two numbers?

STATS IN YOUR WORLD

HE SAID, SHE SAID

If you need data about body weight, should you ask the individuals in your sample or should you take actual measurements? When *asked* their weight, almost all women say they weigh less than they really do. Heavier men also underreport their weight—but lighter men claim to weigh more than the scale shows. We leave you to ponder the psychology of the two sexes. Just remember that "say so" is no substitute for measuring.

6.11 A survey of 100,000 physicians. In 2010, the Physicians Foundation conducted a survey of physicians' attitudes about health care reform, calling the report "a survey of 100,000 physicians." The survey was sent to 100,000 randomly selected physicians practicing in the United States, 40,000 via post office mail and 60,000 via email. A total of 2379 completed surveys were received.[24]

a. State carefully what population is sampled in this survey and what is the sample size. Could you draw conclusions from this study about all physicians practicing in the United States?
b. What is the rate of nonresponse for this survey? How might this affect the credibility of the survey results?
c. Why is it misleading to call the report "a survey of 100,000 physicians"?

6.12 Physician-assisted suicide, continued. Exercise 6.2 described a Gallup study of public opinion toward the topic of physician-assisted suicide. The study used two different versions of the main question. Gallup found that 70% of those who heard version A of the question answered "should be allowed," compared with only 51% of those who heard version B of the question. What is the most likely explanation for this stark difference in opinion? Explain your answer.

COHORTS AND CASE-CONTROL STUDIES

Case-control studies Some biological traits are too rare to study effectively with a sample survey. For instance, a trait found in about 1 in 1000 individuals would require a random sample of 10,000 individuals to observe perhaps about 10 individuals with the trait of interest. It would be much simpler to select two random samples. In epidemiology, this design is often called *case-control:* A random sample of individuals with a condition (the cases) is compared with a random sample of individuals without the condition (the controls).

CASE-CONTROL STUDY

In a **case-control** observational study, case-subjects are selected based on a defined outcome, and a control group of subjects is selected separately to serve as a baseline with which the case group is compared.

EXAMPLE 6.11 A pertussis epidemic in California

Pertussis, also known as whooping cough, is a highly contagious respiratory disease that is characterized by uncontrollable, violent coughing and affects mostly children. California has recently seen a resurgence of cases of pertussis, despite the existence of a vaccine. Researchers compared vaccination records among case children diagnosed with pertussis during the 2010 California epidemic and control children who had no such diagnosis at that time.[25]

Monkeybusiness Images/AGE Fotostock

The cases were a random sample of 682 medical records of California children ages 4 to 10 who had been diagnosed with pertussis. The controls were a random sample of 2016 medical records of California children in the same age group who had received care from the same clinicians but were not diagnosed with pertussis. For every case, about three controls were selected from the clinician's appointment log for that day.

Case-control studies start with two samples of individuals selected for their different outcomes and look for differences between the two groups. Most commonly, researchers look for exposure factors in the subjects' pasts that differ. Looking back into the past is called a **retrospective** approach. In the pertussis study, the researchers examined the children's vaccination history.

The pertussis study found that the cases were much more likely than the controls to have not received any pertussis-containing vaccines, to have not received all recommended five doses of the vaccine, or to have had a longer interval of time since their last vaccination. Of course, differences in vaccination rates could potentially be explained by differences in access to health care. However, because the controls had received care from the same medical providers as the cases, we can make a strong case that access to health care was not a confounding variable in this study.

Case-control studies are an efficient way to examine rare outcomes and often give fast results, within the usual limitations of observational studies. The main challenge is identifying a random sample of control individuals as similar to the case subjects as possible. *Another important restriction of case-control observational studies is that they cannot be used to estimate the proportion of individuals in the population with the rare condition.* This limitation occurs because the investigators choose how many cases and how many controls to study. In the pertussis example, the researchers chose to have examined vaccination records for 682 cases and roughly three times as many controls (2016). This most definitely does not mean that pertussis affects 25% of children.

Cohort studies Some observational studies aim to examine the emergence of a specific condition over time and how it relates to a number of variables. *Cohort* studies enlist a homogeneous group of fairly similar individuals and keep track of them over a long period of time. For the most part, they are **prospective** studies that record, at regular intervals, all sorts of new relevant information about the study participants.[26] After some time, individuals who have developed a condition are then compared with the remaining, unaffected individuals. Observational studies that monitor a sample of individuals repeatedly over time are sometimes called **longitudinal** studies.

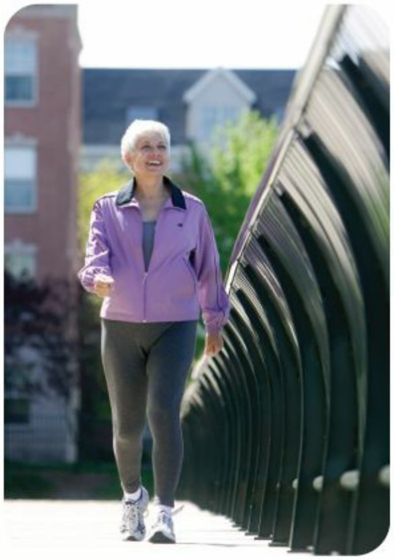
Unlike case-control studies, cohort studies are very costly and better suited to investigating common outcomes such as type 2 diabetes or heart disease. By using the existing medical records of members of a large health care organization to form their cohort, some cohort studies compare members with different conditions in a retrospective approach to identify previous health events that might have influenced later outcomes.

COHORT STUDY

In a **cohort study,** subjects sharing a common demographic characteristic are enrolled and observed at regular intervals over an extended period of time.

EXAMPLE 6.12 The Nurses’ Health Study

The Nurses’ Health Study is one of the largest prospective observational studies designed to examine factors that may affect major chronic diseases in women. Since 1976, the study has followed a cohort of more than 100,000 registered nurses, based on the idea that nurses are able to respond accurately to technically worded medical questionnaires. Every two years, enrolled nurses receive a follow-up questionnaire about diseases and health-related topics such as diet and lifestyle. The response rates to the questionnaires are approximately 90% for each two-year cycle.



Tom Grill/Corbis

In the 2007 newsletter, study investigators reported their findings on age-related memory loss. Approximately 20,000 women age 70 and older had completed telephone interviews every two years to assess their memory with a set of cognitive tests. One of the study findings was that the more women walked during their late 50s and 60s, the better their memory was at age 70 and older. The study was observational because the investigators did not randomly assign different walking regimens. Instead, they observed that women who had walked more during their late 50s and 60s ended up with the better memory scores at age 70 and older.

Cohort studies accumulate enormous amounts of detailed information and can examine the compounded effect of various factors over time. However, they also take a long time to complete and lose subjects over time, creating a potential confounding effect. This is especially true when studying older individuals, who may die before the end of the study. For example, it is possible that the women with the greatest memory loss also had poor overall health and died younger and, therefore, were not included in the analysis.

Overall, though, cohort studies are less prone to confounding than case-control studies because cohorts start with one homogeneous group. This approach has other important advantages. Cohorts can provide information about the relative health risks of different subgroups. They also support incidence calculations, unlike case-control studies, which select subjects based on an existing disease status.

However, like all observational studies, cohort designs cannot establish causation. The observed differences between groups can be attributed either to the groups' differentiating feature or to confounding variables. For example, perhaps the women with the better health were both more capable of walking and less prone to memory loss. Therefore, we cannot unambiguously conclude that walking has a protective effect against memory loss.

Observational studies play an important role in the building of scientific knowledge, but they are vulnerable to bias and confounding. Experiments, by comparison, provide an opportunity for manipulating the environment and confounding variables. In the next chapter we study ways in which experiments can be designed to avoid bias and confounding.

APPLY YOUR KNOWLEDGE

6.13 Antibiotic exposure during infancy. Early life corresponds to the first bacterial colonization of the digestive tract, creating a microbiome that appears to play an important role in human development. What might be the effect of early exposure to antibiotics on children's development? One study followed 11,532 full-term babies all born in the county of Avon (United Kingdom) during 1991 and 1992. Exposures to antibiotics during infancy were recorded, along with recurring measures of body mass over several years. The study found that antibiotic exposure during the first six months of life was associated with increased body mass later on—a pattern not seen with antibiotic exposure occurring after the first six months.[27] What type of observational study is this? What population or populations are represented in the study? What are the explanatory and response variables?

6.14 HPV and oropharyngeal cancer. Molecular evidence suggests a link between the human papillomavirus (HPV) and oropharyngeal cancer. An epidemiologic study enrolled 100 hospitalized patients with newly diagnosed oropharyngeal cancer and 200 hospitalized patients without oropharyngeal cancer. Oral HPV infection was strongly associated with oropharyngeal cancer, with 57 of the 100 cancer patients having an HPV infection, compared with only 14 of the 200 patients without cancer.[28] What type of observational study is this? What population or populations are represented in the study? What are the explanatory and response variables?

Anatomical Travelogue/Science Source

CHAPTER 6 SUMMARY

- We can produce data intended to answer specific questions by conducting **observational studies** or **experiments.** Experiments, unlike observational studies, actively impose some treatment on the subjects of the experiment.
- Observational studies often fail to show that changes in an explanatory variable actually cause changes in a response variable, because the explanatory variable is **confounded** with lurking variables not recorded in the study. Variables are confounded when their effects on a response cannot be distinguished from each other.
- An observational study selects a **sample** from the **population** of all individuals about which we desire information. We base conclusions about the population on data from the sample.
- The **design** of a sample describes the method used to select the sample from the population. **Probability sampling** designs use chance to select a sample.
- The basic probability sample is a **simple random sample (SRS).** An SRS gives every possible sample of a given size the same chance to be chosen. Software or a table of random digits can be used to select an SRS from a known, finite population.
- **Stratified random samples** and **multistage random samples** are more complex forms of probability sampling that are also commonly used.
- Failure to use probability sampling often results in **bias,** or systematic errors in the way the sample represents the population. **Voluntary response samples,** in which the respondents choose themselves, are particularly prone to large bias.
- **Sample surveys** are observational studies of a random sample drawn from the entire population. **Undercoverage** and **nonresponse** are two major survey challenges that can result in substantial bias. In addition, **response bias** and **poorly worded questions** can create misleading conclusions.

- **Case-control** studies select two random samples based on a known characteristic and typically examine **retrospectively** what might have led to the difference. They are well suited to the study of rare conditions. However, finding controls that closely resemble the case-subjects is the most challenging part of this observational design.
- **Cohort** studies enroll subjects who share a common demographic factor at one point in time and typically follow them **prospectively** over time, with data collected at regular intervals. These observational studies are very comprehensive but also expensive and may not be adequate to study rare outcomes.

THIS CHAPTER IN CONTEXT

The methods introduced in Chapters 1 through 5 can be used to describe data regardless of how the data were obtained. However, if we want to reason from data to give answers to specific questions or to draw conclusions about the larger population, then the method that was used to collect the data is important. Sampling is one way to collect data, but it does not guarantee that we can draw meaningful conclusions. Biased sampling methods, such as convenience sampling and voluntary response samples, produce data that can be misleading, resulting in incorrect conclusions. The use of impersonal chance to select the sample is key to a useful sampling process. Probability sampling avoids bias and produces data that can lead to valid conclusions regarding the population. Even with perfect sampling methods, there is still sample-to-sample variation; we will begin our study of the connection between sampling variation and drawing conclusions (statistical inference) in Chapter 13.

Even when we use a simple random sample, our conclusions can be weakened by practical difficulties in obtaining a representative sample, such as undercoverage and nonresponse, and difficulties in obtaining reliable data, such as poor wording of questions and response bias. Careful attention must be given to all aspects of the sampling process to ensure that the conclusions we reach are valid. However, even well-designed observational studies cannot readily be used to determine causality from an observed association (as also discussed in Chapter 4). In Chapter 7 we will examine various types of experimental designs and consider what we can and cannot conclude from them.

CHECK YOUR SKILLS

6.15 The *New England Journal of Medicine* (*NEJM*) posted an opt-in poll on its website, next to an editorial about regulation of sugar-sweetened beverages. The poll asked, “Do you support government regulation of sugar-sweetened beverages?” You just needed to click on a response (yes or no) to become part of the sample. The poll stayed open for several weeks in October 2012. Of the 1290 votes cast, 864 were “yes” responses. You can conclude that

a. approximately two-thirds of Americans support government regulation of sugar-sweetened beverages.
b. approximately two-thirds of *NEJM* readers support government regulation of sugar-sweetened beverages.
c. the poll uses voluntary response, so the results tell us little about any particular population.

6.16 What population is represented by the opt-in poll in Exercise 6.15?

a. All Americans with access to the internet
b. All internet users who visit the *NEJM* website
c. Only those internet users who chose to participate in this particular opt-in poll

6.17 Researchers investigated material-need insecurities among adults diagnosed with diabetes. They found that, in a random sample of 411 adult diabetic patients treated in medical centers all over Massachusetts, 28% reported cost-related medication underuse.[29] The population represented by this study is

a. all adults living in the United States.
b. all adults diagnosed with diabetes living in the United States.
c. all adults diagnosed with diabetes treated in Massachusetts.

6.18 The researchers in Exercise 6.17 had initially selected an additional 113 patients, but these individuals refused to participate in the study. What is the rate of nonresponse for this study?

a. 22%
b. 27%
c. 73%

6.19 The study described in Exercise 6.17 is

a. an observational study with a voluntary response sample.
b. an observational study with a probability sample.
c. a randomized experiment.

6.20 People who experience online harassment either ignore it or they respond via postings or by involving authorities. Is one method more effective than the other? A 2014 Pew Research Center report, based on a random sample of American adults, found that 412 of the 549 respondents who had chosen to ignore online harassment thought that it was an effective way to deal with the issue. Of the 368 respondents who had chosen to respond to online harassment, 305 thought that this approach had been effective. This study is

a. an observational study with a probability sample.
b. a case-control observational study.
c. a randomized experiment.

6.21 A Gallup poll of Americans' smoking behavior and attitudes was conducted in July 2012. A total of 1014 American adults were randomly selected and interviewed by phone. Of the 166 current smokers in the sample, only 1% said that they smoke more than one pack of cigarettes per day, a historical low. In its report Gallup states, "It is possible that the decline in reports of smoking is the result of respondents' awareness that smoking is socially undesirable. Therefore, respondents may aim to present themselves in the best possible light to the interviewer and underestimate the amount they truly smoke." To which type of bias is Gallup referring in this statement?

a. Nonresponse bias
b. Response bias
c. Interviewer bias

6.22 A study enrolled 517 children age 2 to 5 years with a diagnosis of autism spectrum disorder and 315 control children in the same age group but without such a diagnosis. Information was collected about metabolic conditions of the mothers at the time of their pregnancy, such as type 2 diabetes, hypertension, and obesity. Metabolic conditions during pregnancy tended to be more frequent among the mothers of children with autism spectrum disorder than among mothers of control children.[30] This is

a. an observational study with a cohort design.
b. an observational study with a case-control design.
c. an experiment.

6.23 From 1986 to 2000, a study followed a large number of male health professionals age 40 to 75. Each respondent filled out a lifestyle questionnaire every two years. The 22,086 subjects who had reported having good erectile function in 1986 were included in an analysis of risk factors for the development of erectile dysfunction. The analysis revealed that "obesity and smoking were positively associated, and physical activity was inversely associated, with the risk of erectile dysfunction developing."[31] This is

a. an observational study with a cohort design.
b. an observational study with a case-control design.
c. an experiment.

6.24 How strong is the evidence cited in the Excercise 6.23 that physical activity may lower the risk of erectile dysfunction?

a. Quite strong because it comes from an experiment
b. Quite strong because it comes from a large random sample
c. Weak because physical activity is confounded with many other variables

CHAPTER 6 EXERCISES

6.25 “How university students sleep.” This is the title of a report published by Jawbone, a manufacturer of wearable devices used to monitor sleep, movement, and food intake. The report states: “Leveraging the sleep tracking capabilities of UP by Jawbone, we can take an unprecedented look at how tens of thousands of college students sleep across the country at over 100 universities, totaling 1.4 million nights of sleep.” The company’s data showed an average sleep time of 7.0 hours for week nights and 7.4 hours for weekend nights.[32]

a. What type of sample do these data come from? What is the population represented by this sample?
b. What can we conclude from the stated findings?

6.26 Safety of anesthetics. The National Halothane Study was a major investigation of the safety of anesthetics used in surgery. Records of more than 850,000 operations performed in 34 major hospitals showed the following death rates for four common anesthetics:[33]

Anesthetic	A	B	C	D
Death rate	1.7%	1.7%	3.4%	1.9%

There is a clear association between the anesthetic used and the death rate of patients. Anesthetic C appears dangerous.

a. Explain why we call the National Halothane Study an observational study rather than an experiment, even though it compared the results of using different anesthetics in actual surgery.
b. When the study looked at other variables that are confounded with a doctor’s choice of anesthetic, it found that anesthetic C was not causing extra deaths. Suggest important lurking variables that are confounded with which anesthetic a patient receives.

6.27 Sampling Amazon forests. Stratified samples are widely used to study large areas of forest. Based on satellite images, a forest area in the Amazon basin is divided into 14 types. Foresters studied the four most commercially valuable types: alluvial climax forests of quality levels 1, 2, and 3, and mature secondary forest. They divided the area of each type into large parcels, chose parcels of each type at random, and counted tree species in a 20- by 25-meter rectangle randomly placed within each parcel selected. Here is some detail:

Forest type	Total parcels	Sample size
Climax 1	36	4
Climax 2	72	7
Climax 3	31	3
Secondary	42	4

The researchers chose the stratified sample of 18 parcels described in the table. Explain how a stratified random sample differs from a simple random sample. Why do you think the researchers preferred to use a stratified design?

6.28 Telephone area codes. There are approximately 371 active telephone area codes covering Canada, the United States, and some Caribbean regions. (More are created regularly.) You want to choose an SRS of 25 of these area codes for a study of available telephone numbers. Label the codes 001 to 371 (ignore the fact that area codes are numerical sequences themselves), and use the *Simple Random Sample* applet or other software to choose your sample. (If you use Table A, start at line 129 and choose only the first 5 codes in the sample.)

APPLET

6.29 Legalizing marijuana. A January 2010 SurveyUSA poll asked a random sample of adults living in the state of Washington the following question: “State lawmakers are considering making marijuana possession legal. Do you think legalizing marijuana is a good idea or a bad idea?” Of the 500 adults interviewed, 280 said “good idea.” A separate January 2010 SurveyUSA poll asked a random sample of adults living in the San Diego, California, area, “Do you think marijuana should? or should not? be legal when used for recreational purposes?” Of the 500 adults interviewed, 219 answered “should.”

a. What percent of respondents seemed to support the idea of legalizing marijuana in each poll?
b. In what way do these two surveys differ? Explain how these differences might explain the apparent difference of opinion.

6.30 Mischievous responders. A number of studies gather data about adolescents using self-administered questionnaires (SAQ), rather than in-person interviews, to provide anonymity to the participants. Researchers investigated the truthfulness of SAQ answers by adding either reference questions for which most answers should most likely be a negative (such as “Are you currently in a gang?”) or questions for which answers can be compared to the known distribution in the target population (such as “What is your height?”). Here are some of the findings:[34]

Answered that they are currently in a gang	2% of those who said they are not physically disabled
	28% of those who said they are physically disabled

Provided a height in the top or bottom 2.5%	4% of those who said they are not transgender
	42% of those who said they are transgender

a. Some combinations of answers occur much more often than we could reasonably expect, and the researchers attributed them to “mischievous responders.” (For example, the suspiciously high proportion of gang membership among the physically disabled suggests some mischievous responders.) What type of bias would such responses represent?
b. A lot of the research on adolescents aims to find out who may be at risk of depression or suicide and whether minority groups are at a greater risk. Explain why “mischievous responders” are likely to impact the findings on minority groups the most.

6.31 Monitoring chimpanzee behavior. Jane Goodall dedicated years of her life to the study of chimpanzee behavior in their natural habitat of East Africa. Initially, the chimps would flee at the sight of her, so Goodall had to observe them from a distance with binoculars. But she persisted until the animals eventually became accustomed enough to her presence to ignore her.[35]

a. What kind of bias was she trying to minimize?
b. Explain why this lengthy step is particularly important when observing animal behavior.

6.32 Seat belt use. A study in El Paso, Texas, looked at seat belt use by drivers. Drivers were observed at randomly chosen convenience stores. After they left their cars, they were invited to answer questions that included questions about seat belt use. In all, 75% said they always used seat belts, yet only 61.5% were wearing seat belts when they pulled into the store parking lots.[36] Explain the reason for the bias observed in responses to the survey. Do you expect bias in the same direction in most surveys about seat belt use?

6.33 Wording survey questions. Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?

a. “It is estimated that disposable diapers account for less than 2% of the trash in today’s landfills. In contrast, beverage containers, third-class mail, and yard wastes are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?”
b. “Given the current trend of more home runs and more injuries in baseball today, do you think that steroid use should continue to be banned even though it is not enforced?”
c. “In view of the negative externalities in parent labor force participation and pediatric evidence associating increased group size with morbidity of children in day care, do you support government subsidies for day care programs?”

6.34 How much do people sleep? A *New York Times* article[37] described two different studies of Americans’ sleep patterns. The first one, a 2013 Gallup poll, asked “how much time people ‘usually’ sleep,” and found an average of 6.8 hours per night. The second one, the Labor Department’s 2012 American Time Use Survey, reported an average of 8.7 hours of sleep per night based on a diary of how participants “spent their time on the previous day, for each hour of the day.”

a. What are some obvious differences between the two surveys that could explain findings as different as 6.8 and 8.7 hours of night sleep?
b. The article states that *the surveyed populations are also slightly different (18 and older for Gallup, versus 15 and older for the Labor Department).* Explain briefly how this could contribute to the different findings.

6.35 Write your own bad questions. Write your own examples of bad sample survey questions.

a. Write a biased question designed to get one answer rather than another. Then rewrite it to avoid that bias.
b. Write a question that is confusing, so that it is hard to answer. Then rewrite it in an improved form.

6.36 Depression and income level. Gallup reports a survey of depression based on telephone interviews with 258,141 national adults conducted January through September 2009. The findings show that 17% of adult Americans have been told at some point by a physician or nurse that they have depression. However, when looking only at individuals in the lower income bracket (annual income of less than $24,000), that number goes up to 30%. Gallup concludes that lower-income Americans have a higher rate of clinical depression but cautions against attempts to see a causal relationship between income and depression. Here is what Gallup says:

> *It is important to remember that some variables that appear to be risk factors for depression can also be consequences of a major depressive episode. For instance, chronic depression may cause an inability to work, which may lead to poverty. Variables that appear to be risk factors can also be consequences of variables that occur even further up in a causal chain. For instance, serious health problems may*

simultaneously lead to both depression and poverty.

a. What important limitation of surveys, and observational studies in general, does Gallup describe here?
b. Summarize Gallup's argument in a diagram similar to that of Figure 6.1.

6.37 The Gallup-Healthways Well-Being Index. The polling firm Gallup produces extensive monthly and annual reports on the well-being of the American population, including what fraction of the population exercises regularly, has access to health care, or had the flu in any given month. Here are some quotes taken from the 2016 web page describing the survey methodology:

a. *How are interviews conducted for the Gallup-Healthways Well-Being Index? Gallup interviews U.S. adults aged 18 and older living in all 50 states and the District of Columbia using a dual-frame design, which includes both landline and cellphone numbers. Gallup samples landline and cellphone numbers using random-digit-dial methods. Gallup purchases samples for this study from Survey Sampling International (SSI). Gallup chooses landline respondents at random within each household based on which member had the most recent birthday. Each sample of national adults includes a minimum quota of 60% cellphone respondents and 40% landline respondents, with additional minimum quotas by time zone within region. Gallup conducts interviews in Spanish for respondents who are primarily Spanish-speaking.* What is the population that this survey aims to describe? Explain how including both landline telephones and cell phones helps reach the target population. Why do you think that Gallup selects the adult with the most recent birthday in a household, rather than the first one to answer the phone?
b. *Are Gallup-Healthways Well-Being Index samples weighted? Yes, Gallup weights samples to correct for unequal selection probability, nonresponse, and double coverage of landline and cellphone users in the two sampling frames. Gallup also weights its final samples to match the U.S. population according to gender, age, race, Hispanic ethnicity, education, region, population density, and phone status*

(cellphone only, landline only, both, and cellphone mostly). Explain why Gallup weights (adjusts) the answers of its survey respondents. Name one type of bias that cannot be helped with weighting answers.

6.38 Bone loss by nursing mothers. Breast-feeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, resulting in reduced bone mineral density. Researchers compared 47 breast-feeding women with 22 women of similar age who were neither pregnant nor lactating. They measured the percent change in the mineral content of the women's spines over three months.[38]

a. What type of observational study is this? What two populations did the investigators want to compare?
b. Give one possible confounding variable for the relationship between breast-feeding and change in bone mineral content.

6.39 Coffee and depression. A study examined coffee consumption between 1980 and 2004 for 50,739 American nurses from the Nurses' Health Study who were free of depressive symptoms at baseline in 1996. Ten years later, 2607 nurses had reported a diagnosis of depression and antidepressant use at some point. The study found that depression risk decreased with greater caffeinated coffee consumption (but not with decaffeinated coffee consumption).[39]

a. What type of observational study is this? Explain your answer.
b. According to the investigators, "this study cannot prove that caffeine or caffeinated coffee reduces the risk of depression but only suggests the possibility of such a protective effect." Explain why this is the case.
c. What would be a possible confounding variable for this study?

6.40 Brain size and autism. Is autism marked by different brain growth patterns in early life? One study looked at the brain sizes of 30 autistic boys and 12 nonautistic boys who had all received an MRI scan while toddlers. The autistic boys had received the MRI as part of a diagnostic for behavioral concerns, and the nonautistic boys had been participants in a previous quantitative MRI study reporting age-related changes in head and brain features.[40]

a. What type of observational study is this? Explain your answer.

b. The researchers honestly disclosed the limitations of the study, an important part of the ethical conduct of research. Suggest some potential confounding variables for this study.

6.41 Physical activity and heart failure. The California Men's Health Study enrolled 82,695 middle-aged men with no prior history of heart failure. The time each one spent being sedentary was recorded using regularly administered questionnaires over a 10-year period. The study found that the proportion experiencing a heart failure over the course of the study was higher among the men spending more time being sedentary.[41]

a. What type of observational study is this? Explain your answer.

b. During the course of the study, 3473 men experienced heart failure. How is this information relevant when interpreting the study results?

c. Can we conclude from this study that spending a lot of time being sedentary causes heart failure in middle-aged men? Explain your answer.

6.42 Hormone therapy and thromboembolism. Oral estrogen-replacement therapy (ERT) activates blood coagulation and increases the risk of venous thromboembolism (a traveling blood clot, sometimes fatal) in postmenopausal women. A study compared a random sample of 155 postmenopausal women who had experienced a first thromboembolism and a control random sample of 381 postmenopausal women with no history of thromboembolism. One of the study findings was that 21% of the women in the thromboembolism group were currently using oral ERT, compared with only 7% of the women in the control group.[42]

a. What type of observational study is this?

b. Can you conclude that oral ERT causes the higher rate of thromboembolism? Explain your answer.

6.43 Quitting smoking and risk for type 2 diabetes. Researchers enrolled a group of 10,892 middle-aged adults and studied them over a period of nine years. They found that smokers who quit had a higher risk for diabetes within three years of quitting than either nonsmokers or continuing smokers.[43]

a. What type of observational study is this?
b. Does this show that stopping smoking causes the short-term risk for diabetes to increase? Should a doctor cite this study to tell a middle-aged adult patient who smokes that stopping smoking can *cause* diabetes and advise him or her to continue smoking?
c. What confounding variables might explain these findings?

CHAPTER 7 Designing Experiments

cristinairanzo/Getty Images

IN THIS CHAPTER WE COVER...

- Designing experiments
- Randomized comparative experiments
- Common experimental designs
- Cautions about experimentation
- Ethics in experimentation
- *Discussion: The Tuskegee syphilis study*

How do we *produce data* that can help us answer interesting scientific questions? In Chapter 6 we looked at ways to select samples from a population of interest and their applications for observational studies.

But what if we asked, “Is a new cholesterol-lowering drug more effective than the standard-treatment drug?” Giving the new drug to a group of patients with high cholesterol and measuring the decrease in their cholesterol level after one month of treatment wouldn’t be enough to answer the question completely. We would also need to compare their results to those of another group of patients with high cholesterol given the standard drug. In this chapter we study specifically how to design experiments such as this.

DESIGNING EXPERIMENTS

SUBJECTS, FACTORS, TREATMENTS

The **individuals** (or units) studied in an experiment are often called **subjects** when they are people, and **experimental units** otherwise.

The explanatory variables in an experiment are often called **factors.**

A **treatment** is any specific experimental condition applied to the subjects. If an experiment has more than one factor, a treatment is a combination of specific values of each factor.

EXAMPLE 7.1 Blocked tear ducts

Congenital nasolacrimal duct obstruction, or blocked tear duct, is a fairly common condition in infants, characterized by excessive tear production and mucus discharge in one or both eyes. The condition can be treated immediately in office by surgically probing the blocked tear duct after application of a topical anesthetic. An alternative approach is to defer treatment and continue observation until either the condition resolves itself or medical care becomes clearly necessary.

cristinairanzo/Getty Images

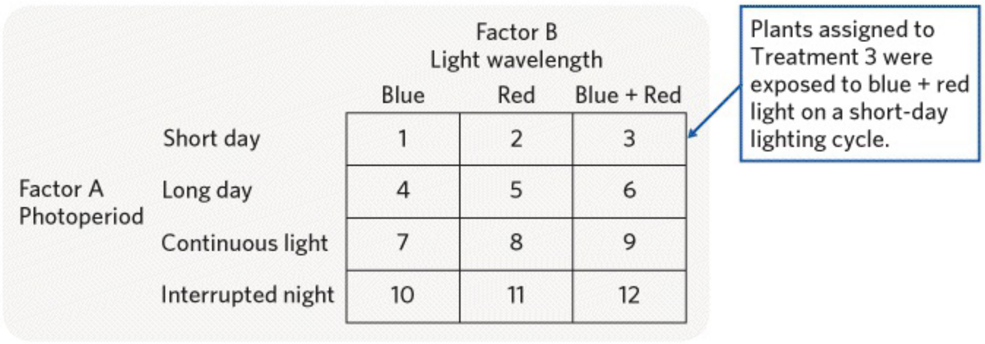
To find out which approach is more effective, researchers assigned infants between six and nine months of age to 2 groups: immediate probing (82 infants) and deferred treatment (81 infants). The infants were examined by a doctor up to eighteen months later to determine if and when the condition had fully resolved. The assignment to one treatment group or the other was random, and the 2 groups did not differ substantially in demographics or initial severity of symptoms. To avoid possible confounding health factors, all infants enrolled in the study were otherwise healthy.[1]

This experiment compares 2 *treatments:* immediate probing and deferred treatment. There is a single *factor:* “treatment type.” The *subjects* are infants diagnosed with blocked tear duct and the main *response variable* is the time of symptom resolution. (The researchers also recorded all related medical care and associated costs.)

EXAMPLE 7.2 Effect of photoperiod and wavelength on flowering

Photoperiod, the relative lengths of light and dark periods in a 24-hour cycle, is a common environmental cue for flowering, allowing plants to bloom during an appropriate time of the year. Chrysanthemums (*Chrysanthemum morifolium*) are known as short-day plants, blooming during times of longer nights. Plant physiologists grew chrysanthemum plants in controlled greenhouses under different combinations of photoperiods (short day, long day, continuous light, and interrupted night) and light wavelengths (blue light, red light, and blue + red light). The plants were kept in these conditions for five weeks and examined regularly to assess whether flowering had occurred during this time.[2]

This experiment has 2 *factors* (explanatory variables): photoperiod, with 4 values (short day, long day, continuous light, and interrupted night), and light wavelength, with 3 values (blue, red, and blue + red). The 12 combinations of both factors form 12 *treatments.* Figure 7.1 shows the layout of the treatments. The *individuals* (or *experimental units*) are chrysanthemum plants, and the *response variable* is flowering, a categorical variable.



		Factor B Light wavelength		
		Blue	Red	Blue + Red
Factor A Photoperiod	Short day	1	2	3
	Long day	4	5	6
	Continuous light	7	8	9
	Interrupted night	10	11	12

Plants assigned to Treatment 3 were exposed to blue + red light on a short-day lighting cycle.

Figure 7.1

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 7.1 The treatments in the experimental design for studying flowering. Combinations of values of the two factors form 12 treatments.

Experiments have clear advantages over observational studies. In an experiment, we can study the effects of the specific treatments in which we are interested. By randomly assigning subjects to treatments, we can avoid confounding. Consider the experiment described in Example 7.1: If the researchers simply compared the medical records of infants treated with immediate probing with those of infants monitored with deferred treatment, it is possible that one group might also be found to have more severe symptoms or to have better access to health care and insurance. Moreover, we can control the environment of the subjects to hold constant factors that are not directly of interest to us. In Example 7.2, for instance, the chrysanthemums were grown in greenhouses where factors such as soil, temperature, and humidity were controlled.

STATS IN YOUR WORLD

WHAT'S NEWS?

Randomized comparative experiments provide the best evidence for medical advances. Do newspapers care? Maybe not. University researchers looked at 1192 articles in medical journals, of which 7% were turned into stories by the two newspapers examined. Of the journal articles, 37% concerned observational studies and 25% described randomized experiments. Among the articles publicized by the newspapers, 58% were observational studies and only 6% were randomized experiments. Conclusion: The newspapers want exciting stories, whether or not the evidence is good.

Another advantage of experiments is that we can study the combined effects of several factors simultaneously. The interaction of several factors can produce effects that could not be predicted from looking at the effect of each factor alone. Perhaps interrupting the night period inhibits flowering in chrysanthemums, but only if the light used during the interruption contains some red light. The two-factor experiment in Example 7.2 will help us find out.

Experiments are the preferred method for examining the effect of one variable on another. By imposing the specific treatment of interest and controlling other influences, we can pin down cause and effect. Statistical designs are often essential for effective experiments, just as they are for sample surveys and observational studies. To see why, let's start with an example of a bad design.

EXAMPLE 7.3 An uncontrolled experiment

Gastric freezing was introduced in the 1960s by a prominent surgeon as a way to relieve ulcer pain. Patients would swallow a deflated balloon that was then filled with a refrigerated liquid. The idea was that the refrigerant would cool the stomach lining, reduce acid production, and reduce pain. The procedures performed showed that gastric freezing did reduce ulcer pain, and the treatment was recommended based on this evidence.[3]

This experiment has a very simple design. A group of subjects (the patients) were exposed to a treatment (gastric freezing), and the outcome (pain reduction) was observed. Here is the design:

$$\text{Subjects} \rightarrow \text{Gastric freezing} \rightarrow \text{Pain reduction}$$

$\text{Subjects} \rightarrow \text{Gastric freezing} \rightarrow \text{Pain reduction}$

Some critical pieces of information are missing from this experiment. What would have happened if we had left the patients alone? Would their pain have abated even without intervention? What if patients were inclined to say they were feeling better (maybe because they expected to get better, or to please the experimenter)? And what if patients genuinely felt better but did so because the care and attention they received made them feel better overall, not because of the cooling of their stomach lining? This is a real possibility, and similar psychological improvements have been documented in many human experiments. We call it the **placebo effect.**

None of these questions can be addressed by the simple experiment. The effect of gastric freezing is confounded with the effect of two lurking variables: spontaneous, natural improvement and psychological improvement due to medical care. Figure 7.2 shows the confounding in picture form. Because of confounding, subject improvement cannot be unequivocally attributed to the effect of gastric freezing.

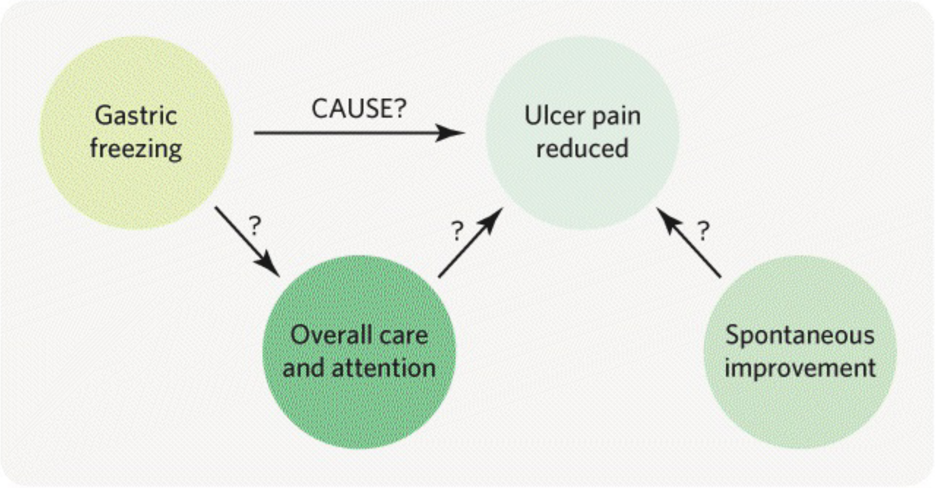


Figure 7.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 7.2 Confounding. We cannot distinguish the effect of the treatment from the effects of lurking variables in the uncontrolled experiment about gastric freezing.

Many experiments performed in a laboratory setting use a design like that of Example 7.3:

$$\text{Subjects} \rightarrow \text{Treatment} \rightarrow \text{Measure response}$$

Subjects → Treatment → Measure response

In the controlled environment of the laboratory, simple designs may work well. However, *field experiments and experiments with human subjects are exposed to more variable conditions and deal with more variable individuals. A simple design often yields worthless results because of confounding with lurking variables.*

CAUTION

APPLY YOUR KNOWLEDGE

7.1 Ginkgo extract and the post-lunch dip. The post-lunch dip is the drop in mental alertness after a midday meal. Does an extract of the leaves of the ginkgo tree reduce the post-lunch dip? A study assigns healthy people age 18 to 40 to take either ginkgo extract or a placebo pill. After lunch, they are asked to read seven pages of random letters and place an X over every *e*. We count the number of misses. What are the individuals, the treatments, and the response variable in this experiment?

7.2 Treating brain cancer. Gliomas are often deadly, aggressive brain cancers. Cannabinoids are thought to inhibit glioma cell growth. A recent study used mice transplanted with glioma cells. The mice were then randomly assigned to one of four cancer treatments: irradiation alone, cannabis alone, irradiation combined with cannabis, or no cancer treatment, for a total of 21 days. The glioma tumor volume (in mm^3) was assessed with brain imaging at the end of the study. Identify the individuals, the factors, and the response variable in this experiment. Use a diagram like that in Figure 7.1 to display the treatments.

RANDOMIZED COMPARATIVE EXPERIMENTS

The remedy for the confounding in Example 7.3 is to do a *comparative experiment* in which some patients undergo the gastric-freezing treatment and other, similar patients undergo a *sham* procedure or no procedure at all. Comparing patients undergoing gastric freezing with untreated patients allows us to control for the effect of natural improvement. However, that alone would not tell us whether a lower pain rate in the *experimental group* (gastric freezing) is due to the benefits of the medical procedure itself or to the psychological impact of undergoing any kind of treatment—the *placebo effect*. Only by comparing the gastric-freezing procedure with a sham procedure would we be able to conclude that gastric freezing itself is responsible for any observed improvement in patient condition.

Most well-designed experiments compare two or more treatments. Part of the design of an experiment is a description of the factors (explanatory variables) and the layout of the treatments, with comparison as the leading principle.

EXPERIMENTAL GROUP, CONTROL, PLACEBO

An **experimental group** is a group of individuals receiving a treatment whose effect we seek to understand.

A **control** group serves as a baseline with which the experimental group is compared.

A **placebo** is a control treatment that is fake (for example, taking a sugar pill) but otherwise indistinguishable from the treatment in the experimental group.

EXAMPLE 7.4 A controlled experiment

The simple experiment outlined in Example 7.3 was followed up many years later with a randomized, controlled experiment in which patients undergoing gastric freezing were compared with similar patients undergoing the same procedure except that the liquid pumped into the balloon was not refrigerated. This second group was a placebo group.

> The comparative study found that the condition of 28 of the 82 patients who received the gastric-freezing treatment improved, while the condition of 30 of the 78 patients in the placebo group improved.
>
> This comparative experiment demonstrated that cooling the stomach lining was not more effective in reducing ulcer pain than a sham procedure. The treatment was then abandoned.

The *placebo effect* is documented in human experimentation, in contexts as diverse as the study of asthma or high blood pressure. The effect is particularly strong when the response variable is pain. The mechanisms of this mind–body effect are unknown but are not without biological foundation. A neural response to the placebo effect has been documented with neuroimaging techniques as early in the sensory process as the spinal cord.[4] We also know that the nervous system interacts with the immune system and that depression, for instance, suppresses the immune response.

Like experiments, observational studies often compare different groups and may even include a control group that does not exhibit the characteristic of interest. For example, an observational study could follow individuals over time and record diagnoses of lung cancer, comparing a sample of cigarette-smoking adults and a sample of nonsmoking adults. However, in observational studies the groups compared may differ by more characteristics than the ones used to distinguish the groups. For example, the smokers may have poorer health or live in worse environments, making them more susceptible to lung cancer. Controls in an observational study are very informative, but they are not sufficient to conclude about a causal effect.

Thus, comparison alone isn't enough to produce strong conclusions. If the treatments are given to groups that differ markedly when the experiment begins, bias will result. In Example 7.4, if we select all the patients most likely to improve to make up the gastric-freezing treatment group, we will see better results in the treatment group even if the procedure itself makes no contribution to the improvement. Any personal choice would bias our results in the same way that volunteers bias the results of opt-in polls. The solution to the problem of bias is the same for experiments and for survey samples: *randomization*. That is, we must use impersonal chance to select the groups.

RANDOMIZED COMPARATIVE EXPERIMENT

An experiment that uses both comparison of two or more treatments and chance assignment of subjects to treatments is a **randomized comparative experiment.**

EXAMPLE 7.5 Acupuncture for migraines

Acupuncture is an ancient traditional Chinese medicine that uses very thin needles inserted at specific body points to relieve pain and other conditions. Researchers investigated whether acupuncture can help prevent migraine attacks. Patients suffering from migraines were randomly assigned to either acupuncture, sham acupuncture (needles inserted at nontraditional anatomic points), or a waiting list for acupuncture. The sham acupuncture acted as a placebo. The waiting-list group acted as a control (no medical treatment). The patients were asked to record headache days in diaries for 12 weeks.[5] Figure 7.3 outlines the design of this study in graphical form.

APPLET

The selection procedure is exactly the same as it is for sampling: label and select. **Step 1. Label** the 302 patients as 001 to 302. **Step 2. Select.** Use software, such as the *Simple Random Sample* applet, or use a table of random digits, such as Table A at the back of the book (reading successive three-digit groups). The first 145 labels encountered select the acupuncture group. Ignore repeated labels and groups of digits not used as labels. After the first group is selected, the next 81 labels encountered select the sham acupuncture group. The remaining 76 labels make up the waiting-list group.

SuperStock/Alamy Stock Photo

The design in Figure 7.3 is *comparative* because it compares three treatments (acupuncture, sham acupuncture, and waiting list). It is *randomized* because the subjects are assigned to the treatments by chance. The “flowchart” outline presents all the essentials: randomization, the sizes of the groups and which treatment they receive, and the response variable. In this study, the researchers chose to place about half the patients in the acupuncture group and to split the other half between the two control treatments (sham and waiting list).

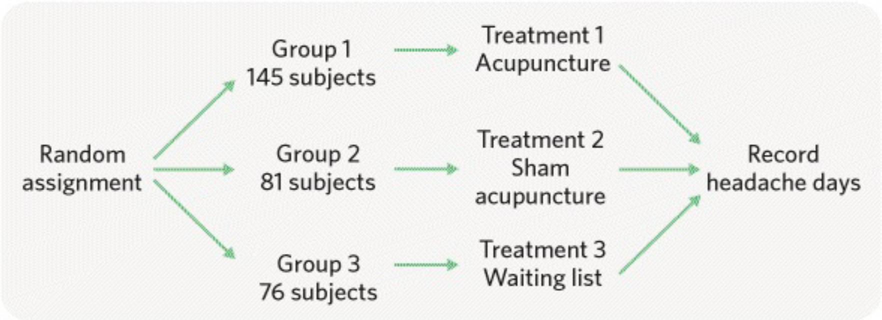

Random assignment
Group 1 145 subjects
Group 2 81 subjects
Group 3 76 subjects
Treatment 1 Acupuncture
Treatment 2 Sham acupuncture
Treatment 3 Waiting list
Record headache days

Figure 7.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 7.3 Outline of a randomized comparative experiment to compare the effect of acupuncture, sham acupuncture, and placebo on headaches.

Randomized comparative experiments are designed to provide good evidence that differences in the treatments actually *cause* the differences we see in the response. The logic is as follows:

APPLET

- Random assignment of subjects forms groups that should be similar in all respects before the treatments are applied. Exercise 7.53 uses the *Simple Random Sample* applet to demonstrate this process. The random assignment also allows us to use probability laws to conduct statistical inference on the data collected, as we will see in later chapters.
- Comparative design ensures that influences other than the experimental treatments operate equally on all groups.
- Therefore, differences in average response must be due either to the treatments or to the play of chance in the random assignment of subjects to the treatments.

This last statement deserves more thought. *Even with an excellent experimental design, we can never say that any difference between two groups is entirely due to differences in the treatments applied.* Variables, by definition, vary from individual to individual. We would not trust an experiment like that of Example 7.5 if it had just one patient in each group because the results would depend too much on which group got lucky and received a patient with less frequent headaches, for example. If we assign many subjects to each group, however, the effects of chance will average out, and there will be little difference in the *average* responses in the three groups unless the treatments themselves cause a difference. "Use enough subjects to reduce chance variation" is the third big idea of statistical design of experiments.

PRINCIPLES OF EXPERIMENTAL DESIGN

The basic principles of statistical design of experiments are:

1. Control the effects of lurking variables on the response, most simply by comparing two or more treatments.

2. Randomize—use impersonal chance to assign subjects to treatments.

3. Use enough subjects in each group to reduce chance variation in the results.

We hope to see a difference in the responses so large that it is unlikely to happen just because of chance variation. As we will see in Part II of this book, we can use the laws of probability, which give a mathematical description of chance behavior, to determine whether a treatment's effect is larger than we would expect to see if only chance assignment to groups were operating. If it is, we call this effect *statistically significant.*

STATISTICAL SIGNIFICANCE

An observed effect so large that it would rarely occur by chance is called **statistically significant.**

If we observe statistically significant differences among the groups in a randomized comparative experiment, we have good evidence that the treatments actually caused these differences. You will often see the phrase “statistically significant” in scientific reports. The great advantage of randomized comparative experiments is that they can produce data that give good evidence for a cause-and-effect relationship between the explanatory and response variables. We know that in general a strong *observed* association does not imply causation. In contrast, a statistically significant association in data from a well-designed *experiment* does imply causation.

APPLY YOUR KNOWLEDGE

7.3 Investigating health claims. In the United States, marketing “healthy” food and supplements does not require demonstrated health benefits, though any *specific* health claim must be backed by scientific evidence. False and unsubstantiated advertising claims are investigated by the Federal Trace Commission (FTC).

The makers of PomWonderful juice and POMx supplements had made a number of advertising claims such as “SUPER HEALTH POWERS! 100% PURE POMEGRANATE JUICE. Backed by $25 million in medical research. Proven to fight for cardiovascular, prostate and erectile health.” After conducting an investigation, the FTC ordered “POM’s future disease treatment and prevention claims to be supported by at least one randomized, well-controlled human clinical trial.”[6]

a. A clinical trial is an experiment using human subjects for health-related research purposes. Explain why the FTC insists that they should be “randomized.”

b. Explain the importance of a clinical trial being “well-controlled” in the context of POM products.

c. The FTC also prohibited POM makers from misrepresenting scientific evidence, stating, for instance, that “the erectile dysfunction claims are false and unsubstantiated because the study on which the company relied did not show that POM Juice was any more effective than a placebo.” Explain the reasoning behind this argument.

7.4 Adolescent obesity. Adolescent obesity is a serious health risk affecting more than 5 million young people in the United States alone. Laparoscopic adjustable gastric banding has the potential to provide a safe and effective treatment. Fifty adolescents between 14 and 18 years old with a body mass index higher than 35 were recruited from the Melbourne, Australia, community for the study. Twenty-five were randomly selected to undergo gastric banding, and the remaining 25 were assigned to a supervised lifestyle intervention program involving diet, exercise, and behavior modification. All subjects were followed for two years and their weight loss was recorded.[7]

a. Outline the design of this experiment, following the model of Figure 7.3. What is the response variable?

b. Carry out the random assignment of 25 adolescents to the gastric-banding group, using the *Simple Random Sample* applet, other software, or Table A (starting at line 130, and starting your labels at 1). The remaining 25 subjects will make up the lifestyle intervention group.

7.5 More rain for California? The changing climate will probably bring more rain to California, but we don’t know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Kenwyn Suttle of the University of California at Berkeley and his coworkers carried out a randomized controlled experiment to study the effects of more rain in either season. They randomly assigned plots of open grassland to 3 treatments: added water equal to 20% of annual rainfall either during January to March (winter) or during April to June (spring), and no added water (control). Thirty-six circular plots of area 70 square meters were available, of which 18 were used for this study. One response variable was total plant biomass, in grams per square meter, produced in a plot over a year.[8]

a. Outline the design of this experiment, following the model of Figure 7.3.

b. Number all 36 plots and choose 6 at random for each of the 3 treatments. Be sure to explain how you did the random selection.

COMMON EXPERIMENTAL DESIGNS

Completely randomized designs The simplest experimental design assigns the individuals (subjects or experimental units) to the treatments completely at random. We call such designs *completely randomized*. Example 7.5, about acupuncture for migraines, describes an experiment with a completely randomized design, and Figure 7.3 illustrates the randomization process. *Note that it is not necessary for a completely randomized design to assign the same number of individuals to each treatment.* It might be advantageous or more practical to assign a larger number of individuals to one particular treatment. This is the case for the acupuncture group in Example 7.5.

COMPLETELY RANDOMIZED DESIGN

In a **completely randomized** experimental design, all the individuals are allocated at random among all the treatments.

Completely randomized designs can compare any number of treatments. They can also have more than one factor. The chrysanthemum-flowering experiment of Example 7.2 has two factors: photoperiod and light wavelength. Their combinations form the 12 treatments outlined in Figure 7.1. A completely randomized design assigns individual chrysanthemum plants at random to these 12 treatments. Once the layout of treatments is set, the randomization needed for a completely randomized design is tedious but straightforward, as illustrated in Example 7.5.

Completely randomized designs are the simplest statistical designs for experiments. They illustrate clearly the principles of comparison, randomization, and adequate number of subjects. However, experiments sometimes need additional sophistication—for example, to deal with substantial individual-to-individual variability or potentially confounding variables.

Block designs One such design is the *block design*, in which individuals sharing the same characteristic are pooled, forming a *block.* A preexisting characteristic is used to form the basis of the blocks; it represents a feature of the individuals rather than an imposed condition or treatment.

BLOCK DESIGN

A **block** is a group of individuals that are known before the experiment to be similar in some way that is expected to affect the response to the treatments.

In a **block design,** the random assignment of individuals to treatments is carried out separately within each block.

Block designs can have blocks of any size. Blocks are another form of *control.* They control for the effects of some outside variables by bringing those variables into the experiment to form the blocks. Here are some typical examples of block designs.

EXAMPLE 7.6 Smoking exposure and asthma care

An estimated 9.5% of children in the United States suffer from asthma. Children with persistent asthma should take daily preventive anti-inflammatory medication. Unfortunately, this recommendation is not always followed. A study compared the effect of the usual at-home asthma care and a school-based asthma care program (in which school nurses administer treatment) on the asthma symptoms of elementary school children. Environmental tobacco smoke can trigger and worsen asthma symptoms, so the experiment used home exposure to tobacco smoke as a blocking variable.[9] Two separate randomizations were done: One assigned the children exposed to smoke to the two treatments, and the other assigned the children not exposed to smoke to the two treatments. Figure 7.4 outlines the design of this experiment. Note that no randomization was involved in making up the blocks. Instead, the groups of subjects differed in some way (home smoke exposure, in this case) that existed before the experiment began.

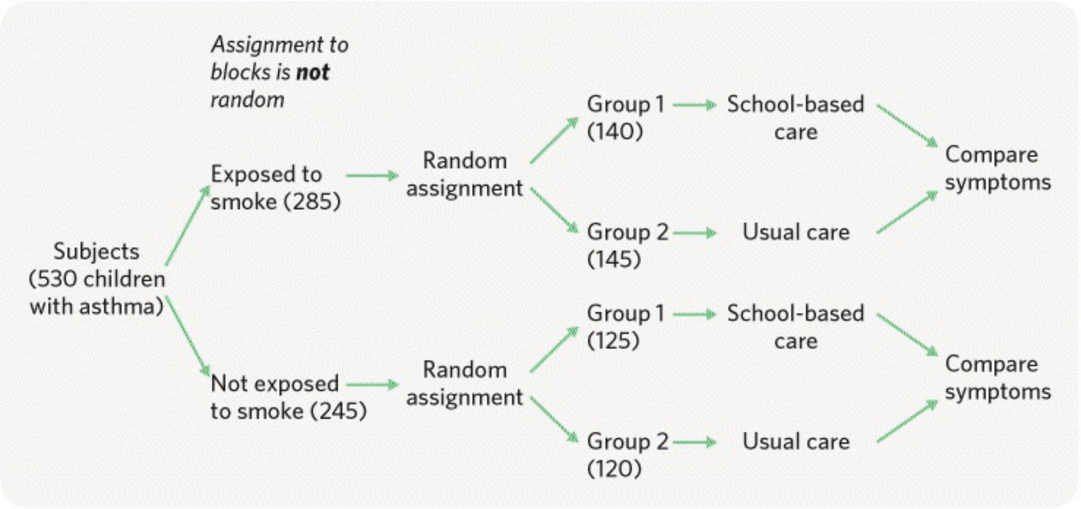

Assignment to blocks is **not** random
Subjects (530 children with asthma)
Exposed to smoke (285)
Random assignment
Group 1 (140)
School-based care
Group 2 (145)
Usual care
Compare symptoms
Not exposed to smoke (245)
Random assignment
Group 1 (125)
School-based care
Group 2 (120)
Usual care
Compare symptoms

Figure 7.4

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 7.4 Outline of a block design comparing two asthma care options for asthmatic children who are or are not exposed to smoke at home.

EXAMPLE 7.7 Soil differences

The soil type and fertility of farmland differ by location. To account for these differences, experiments on plant growth often divide the available planting area into smaller fields, which make up the blocks. A test of the effect of tillage type (two types) and pesticide application (three application schedules) on soybean yields used small fields as blocks. Each block was further divided into six plots, and the six treatments were randomly assigned to plots separately within each block, as illustrated in Figure 7.5. Field location, used to create the blocks, was a preexisting feature of the experimental units (the plots).

IIC/Axiom/Getty Images

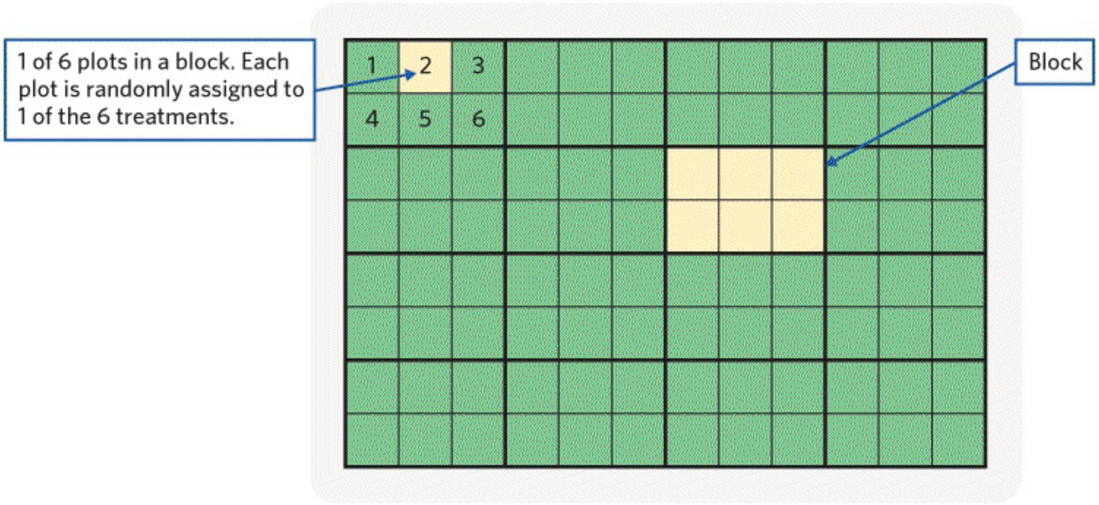

1 of 6 plots in a block. Each plot is randomly assigned to 1 of the 6 treatments.
1
2
3
4
5
6
Block

Figure 7.5

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 7.5 Outline of a block design. The blocks are land areas subdivided into six planting plots. The treatments are six combinations of tillage type and pesticide application.

In both examples, blocking allowed the experimenters to include in their analysis an additional variable (smoke exposure in Example 7.6 and field location in Example 7.7) that was not an imposed treatment in the experiment. Blocks allow us to draw separate conclusions about each block and to assess whether the response to the treatments (such as the two asthma care options in Example 7.6) may depend on the block variable.

The idea of blocking is an important additional principle of statistical design of experiments. A wise experimenter will form blocks based on the most important unavoidable sources of variability among the subjects or experimental units. Randomization will then average out the effects of the remaining variation and allow an unbiased comparison of the treatments.

Matched pairs designs Another experimental design is especially useful when substantial differences between individuals exist or are suspected. A *matched pairs design* uses a form of blocking to compare just two treatments. With this type of design, the experimenter chooses pairs of subjects that are as closely matched as possible: These can be individuals of the same sex, age, weight, and stature, for example, or they can be genetically related individuals such as twins or animals born in the same litter. The experimenter then uses chance to decide which subject in a pair gets one specific treatment. The other subject in that pair gets the other treatment. That is, the random assignment of subjects to treatments is done within each matched pair, not for all subjects at once. Alternatively, sometimes each "pair" in a matched pairs design consists of just one subject, who gets both treatments one after the other. This approach is sometimes called a **crossover design.** In such a case, each subject makes up one pair of experimental results. *Because the* order *of the treatments can influence the subject's response, we randomize the* order *for each subject.*

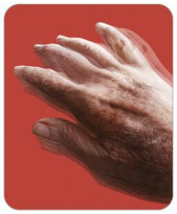
MATCHED PAIRS

A **matched pairs design** compares exactly two treatments, either by using a series of individuals that are closely matched two by two or by using each individual twice.

Matched pairs designs require that the assignment of the two treatments within each "pair" be **randomized** to avoid a systematic bias.

EXAMPLE 7.8 Essential tremor

Essential tremor is a neurological movement disorder characterized by involuntary rhythmic movement that typically interferes with the full use of the arms and hands. The most common movement disorder, it differs from Parkinson's disease in that it is typically action related. Bioengineers designed a noninvasive handheld device using active cancellation of tremor (ACT) technology to stabilize tremor-induced motion.

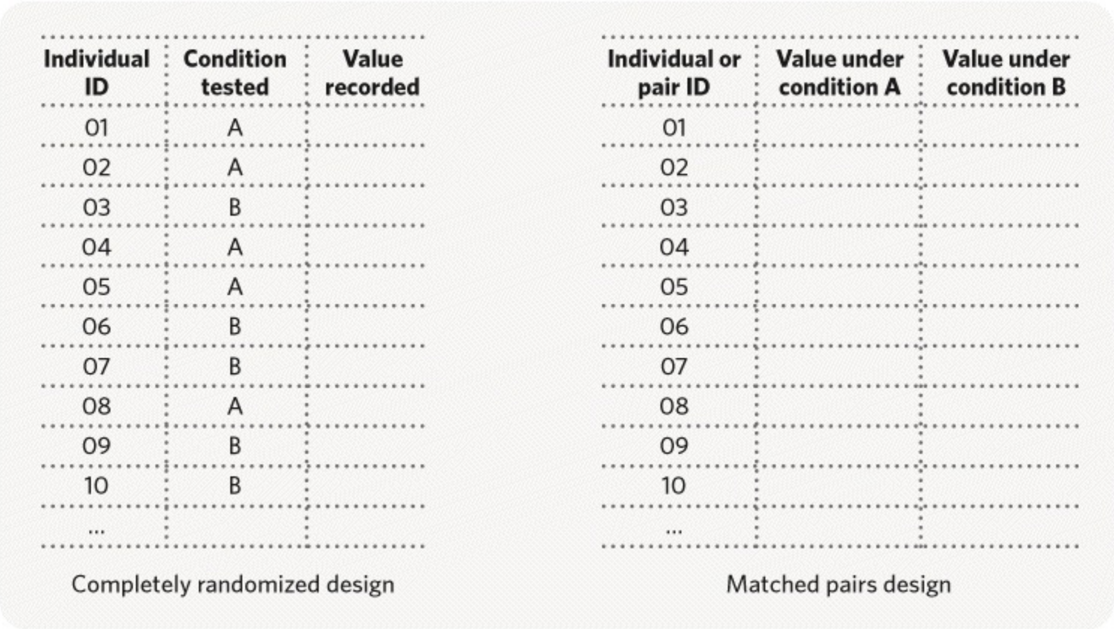


Gwen Shockey/Getty Images

To test the efficacy of the ACT device, the researchers recruited a sample of 15 subjects suffering from essential tremor. The subjects performed a feeding task (bringing a full spoon to their mouth) with the device attached to their dominant hand. The device was turned both on and off in random order during the task, without letting the subjects know when the switch occurred. Tremor amplitude was measured (in centimeters) with a precision accelerometer under both conditions.[10]

This study uses a *matched pairs design* in which pairs of measurements are obtained for each subject. The order of the on and off conditions was randomized in case tremor amplitude changed during the course of the task. Not informing the subjects when the device was switched on or off meant that they could not anticipate the switch and change how they performed the task.

In a *completely randomized design*, the 15 subjects would have been assigned at random to two groups (ACT on and ACT off), so that only one tremor amplitude measurement would have been obtained for each subject. Some patients, however, have more severe conditions characterized by greater tremor amplitudes. The completely randomized design relies on chance to distribute the more severe cases roughly evenly between the two groups. The matched pairs design, in contrast, compares each subject’s tremor amplitude with the device on and with the device off. This makes it easier to see the effect of the ACT device on tremor amplitude despite differences in tremor severity. Figure 7.6 illustrates how data organization differs when data are collected using a completely randomized design (left) versus a matched pairs design (right).



Individual ID	Condition tested	Value recorded
01	A	
02	A	
03	B	
04	A	
05	A	
06	B	
07	B	
08	A	
09	B	
10	B	
...		

Completely randomized design

Individual or pair ID	Value under condition A	Value under condition B
01		
02		
03		
04		
05		
06		
07		
08		
09		
10		
...		

Matched pairs design

Figure 7.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 7.6 Data collection for an experiment with two conditions, A and B, under a completely randomized design (left) versus a matched pairs design (right).

Fields that study mainly human subjects, such as psychology and sociology, often reserve the term “matched pairs” to actual pairs of similar but different individuals. When the same individuals are used for both treatments, the term **repeated measures** is often used instead. One added bonus of using the term “repeated measures” is that it also applies to experiments in which each subject is exposed to more than two treatments.

EXAMPLE 7.9 Chocolate milk for athletes

Chocolate milk is gaining in popularity among athletes as a sports drink because it contains proteins in addition to the carbohydrates (sugar) and minerals found in commercial sports drinks such as Gatorade. Researchers wanted to see if drinking chocolate milk would result in higher endurance among athletes than drinking Gatorade or mineral-enriched water.

SHOW US YOUR
CHOCOLATE
MILK
MUSCLE!

Wisconsin Milk Marketing Board, Inc.

> Athletes vary substantially in their physical abilities. They vary so much that individual variations across athletes might be more substantial than the effect of one drink over another. Recognizing this fact, the researchers asked trained male cyclists to cycle to exhaustion in 3 different weeks: once drinking chocolate milk, once drinking Gatorade, and once drinking mineral-enriched water, in a randomized order. That is, each athlete was subjected to all 3 treatments, in a repeated-measures design.[11]

Matched pairs and repeated-measures designs use the principles of comparison of treatments and randomization. However, the randomization is not complete—we do not randomly assign all the subjects at once to the treatments. Instead, we randomize only within each matched pair or, when using the same individuals for different treatments, we randomize the order of the treatments for each individual. This differs from a completely randomized design, which produces groups that are *independent:* That is, the individuals in one group are completely unrelated to the individuals in the other group(s). *Being able to distinguish a matched pairs design from a completely randomized design is critical because it impacts the statistical analysis and conclusions,* as we will see in the later chapters on inference.

Students sometimes confuse the case-control design described in Chapter 6 and the matched pairs design described in this chapter. *The case-control design applies to observational studies (no treatment is imposed), whereas the matched pairs design described in this chapter applies to experiments.* Also, the selection of control individuals in case-control studies does not necessarily follow a strict matching process, and investigators may select more than one control for each case.

Like the design of samples, the design of experiments can get quite complex. Complex designs have both advantages and disadvantages. On the one hand, in a complex design, the causal effect of one variable on another might be better isolated from influences of other factors. On the other hand, the data collection, analysis, and interpretation are likely to become much more complex. In the rest of the book, we will focus on simple completely randomized, matched pairs, and block designs.

APPLY YOUR KNOWLEDGE

7.6 Psoriasis treatment. Psoriasis is an autoimmune disease manifested in chronic skin inflammation. Researchers recruited adult patients with moderate to severe plaque psoriasis and randomly assigned them to one of 4 treatment options: placebo (309 subjects), ustekinumab (300 subjects), low-dose brodalumab (610 subjects), and high-dose brodalumab (612 subjects). Patients were evaluated after twelve weeks of treatment to determine whether their symptoms were still visible or not.[12]

a. What are the individuals, the treatments, and the response variable in this experiment?

b. What is the design of this experiment? Outline this design as in Figure 7.4.

Jodi Jacobson/Getty Images

7.7 Peanut allergy. The timing of peanuts' first introduction in children's diet could affect allergic reactions, so physicians have debated whether to recommend early exposure or avoidance. A study enrolled 640 infants with severe eczema and split them into two groups based on whether they also had a diagnosed peanut allergy. Within each group, the researchers randomly assigned the infants to either completely avoid peanuts or consume peanuts in small amounts regularly until they reach 60 months of age. At the end of the study, an allergy test was conducted to establish which children had a peanut allergy.[13]

a. What are the individuals, the treatments, and the response variable in this experiment?

b. What is the design of this experiment? Outline this design as in Figure 7.4.

c. Of the 640 infants in the study, only 98 had a diagnosed peanut allergy at the beginning of the study. Explain why you think the researchers decided to split the infants into two groups before randomly assigning them to the conditions.

7.8 Bedtime reading. A bit of bedtime reading is a fairly common practice, and researchers examined whether changes in reading technology could impact sleep quality. They recruited a sample of 12 healthy adults to sleep in a sleep lab on two different days. In random order, participants read, for 30 minutes, a print book on one night and a light-emitting e-reader on the other night. Scalp electrodes were used to measure how long (in minutes) it took participants to reach a deep sleep stage. On average, the subjects took 12.3 minutes longer to reach deep sleep when they read with the e-reader than when they read with the book.[14]

a. What are the individuals, the treatments, and the response variable in this experiment?

b. What is the design of this experiment?

c. Sleeping in an unfamiliar environment can create a sleep disturbance called the first night effect. Explain why it is important to select the print-book and e-reader nights in random order for each participant.

7.9 How long did I work? A psychologist wants to know if the difficulty of a task influences our estimate of how long we spend working at it. She designs two sets of mazes that subjects can work through on a computer. One set has easy mazes and the other has hard mazes. Subjects work until told to stop (after six minutes, but subjects do not know this). They are then asked to estimate how long they worked. The psychologist has 30 students available to serve as subjects.

a. Describe the design of a completely randomized experiment to learn the effect of task difficulty on estimated time spent on the task.

b. Describe the design of a matched pairs experiment using the same 30 subjects.
c. Describe the design of a block experiment considering men and women separately.

CAUTIONS ABOUT EXPERIMENTATION

The logic of a randomized comparative experiment depends on our ability to treat all the subjects identically in every way except for the actual treatments being compared. Good experiments therefore require careful attention to details.

EXAMPLE 7.10 Essential tremor, continued

In Example 7.8 we saw how researchers used a matched pairs design to test the efficacy of the ACT device in patients with essential tremor. They concluded that the device significantly reduced spoon tremor, based on accelerometer data and on clinical evaluation by a neurologist. This means that the tremors were so much smaller with the ACT-on condition that chance alone would not likely produce such results.

This finding is not quite enough, however, to convince us that the ACT device is effective. What if the subjects or the neurologist were somehow biased (consciously or not) by the knowledge of the device being activated? Such bias is unlikely in this case because the study was *double-blind:* As the researchers reported, “subjects and the evaluating neurologist were not told whether the device was on or off. The device prototype also does not emit a sound or cause a palpable sensation that can be detected by the user when turned on.” Who turned the device on and off? A person not otherwise involved with the experiment and, therefore, unlikely to influence the results.

DOUBLE-BLIND EXPERIMENTS

In a **double-blind** experiment, neither the subjects nor the people who interact with them know which treatment each subject is receiving.

Double-blinding is obviously necessary whenever the investigator evaluates the experimental outcome. In clinical trials, how a doctor interacts with a patient can also influence the patient's perception of his or her prognosis, thereby creating or reinforcing a placebo effect. Double-blinding controls the influence of these human interactions by preventing potential bias. In many medical studies, only the statistician who does the randomization knows which treatment each patient is receiving.

Blinding is a challenge if subjects are able to differentiate the treatments. Example 7.5 described a study of acupuncture to treat migraines in which real acupuncture was compared with a sham acupuncture treatment. But how would you compare acupuncture with medication while avoiding a placebo effect? One solution, described in the following example, involves a **double-dummy.**

EXAMPLE 7.11 Acupuncture versus medication

A study compared the effect of acupuncture and the established drug flunarizine in preventing migraines among adult subjects suffering from recurring migraines. Patients can obviously tell the difference between needles and pills, but they cannot tell the difference between real and sham acupuncture or between a flunarizine pill and a placebo pill. So the researchers gave all subjects both types of treatment (needles and pills), in a double-dummy design. The subjects were randomly assigned to two groups: One group received real acupuncture and placebo pills for four weeks; the other group received sham acupuncture and actual flunarizine pills for four weeks. All subjects kept diaries to record migraine occurrences and intensity.[15]

In practice, there isn't always a solution for every experimental challenge. Sometimes, the use of a placebo is unethical. Blinding might not always be practical. For instance, the experiment in Example 7.11 was **single-blind** rather than double-blind, because the medical practitioner performing the real and sham acupuncture had to know which treatment each subject received.

Many—perhaps most—experiments have some weaknesses in detail. The environment of an experiment can influence the outcomes in unexpected ways. Although experiments are the gold

CAUTION

standard for obtaining evidence of cause and effect, really convincing evidence usually requires that the study be **replicated** successfully by its investigators as well as by independent investigators in different locations. All researchers are required to keep detailed records of their experiments and must provide these records if other teams have been unsuccessful at replicating the published results.

EXAMPLE 7.12 Experimenter influence

We know that mice and rats can differentiate human experimenters by smell. So, when it comes to study results, does it matter who the experimenter is? Researchers decided to investigate this question with a randomized experiment, using rodent facial grimacing indicative of pain as an observable behavior.[16] Male and female adult mice that had never before participated in a lab experiment were injected with an irritating product in the ankle. Each mouse was then placed next to an experimenter sitting quietly in an empty room, and the mouse's facial expressions were recorded. Each mouse was tested only once, with a randomly assigned experimenter.

The data show that facial grimacing was significantly diminished when the experimenter was male. The same was true when the mice were left in the room with only a T-shirt previously worn by either a male or a female experimenter. When the study was replicated with lab rats, the findings were similar. The researchers concluded that the sex of an experimenter can affect the behavioral response of rodents in a laboratory setting.

CAUTION

Always be very cautious in your interpretation of an experiment if the likely sources of bias have not all been carefully controlled for.

The most serious potential weakness of experiments is **lack of realism.** The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study. Carcinogenicity, for example, is often tested in extreme situations unlikely to reflect realistic human use.

EXAMPLE 7.13 Carcinogenicity of saccharin

The carcinogenicity of a product is its propensity to induce cancer when used, ingested, or simply present in the environment. Saccharin has been used as an artificial sweetener for almost a century. Because of experimental evidence of carcinogenicity in rats, in 1981 saccharin was listed in the U.S. *Report on Carcinogens* as reasonably anticipated to be a human carcinogen.

However, closer examination of the observed urinary bladder cancers in rats indicate that the biological mechanism of this cancer is specific to the rat urinary system and related to saccharin consumption at concentrations unlikely to represent realistic human consumption. Saccharin was officially delisted from the *Report on Carcinogens* in 2005, with the conclusion that "the factors thought to contribute to tumor induction by sodium saccharin in rats would not be expected to occur in humans."[17]

Lack of realism can limit our ability to apply the conclusions of an experiment to the settings of greatest interest. Most experimenters want to generalize their conclusions to some setting wider than that of the actual experiment. For example, experiments in the health sciences aim to produce results that can be generalized to the entire population or to all individuals in the entire population diagnosed with a particular condition. This is very difficult to do when women and minorities are underrepresented or when potential differences between sexes or ethnic groups are not investigated.

EXAMPLE 7.14 Gender and clinical research

Researchers and drug manufacturers have historically avoided testing new drugs on women. Yet these same drugs have been routinely prescribed regardless of gender. Of course, an untested drug could potentially be harmful to the fetus if a woman were to become pregnant. In drug studies, this concern can easily be alleviated by enrolling only women willing to use contraceptives over the timeline of the clinical trial.

The major reasons for omitting women from clinical research are variations in women’s menstrual cycle and hormonal changes over time with menopause or oral contraceptive use. Response to treatment might be influenced by these hormonal fluctuations, confounding the results. Or the treatment itself might interfere with concurrent use of oral contraceptives. Hormonal fluctuations during the menstrual cycle and the use of an oral contraceptive have indeed been shown to affect how drugs are metabolized. This is true, for example, of such common drugs as acetaminophen, aspirin, diazepam, and even caffeine. Some drugs, such as the antibiotic rifampin, also reduce the effectiveness of oral contraceptives.[18] The question is: Is this a good reason to exclude women from clinical trials?

Absolutely not! We know that samples are representative of only the population from which they are taken. Clinical trials based entirely on men ensure drug effectiveness and safety for men alone. Not including women in clinical research does not protect them: It merely delays the assessment of efficacy and side effects until the drug is made available to the public.

Since the mid-1990s, the Food and Drug Administration (FDA) has required that clinical trials for drugs targeting both men and women include participants of both sexes, in proportions reflecting the actual proportions of men and women in the target population.

A disturbing consequence of women’s historical underrepresentation in clinical trials is that medications were approved for use in women without knowing whether they would work at the dosage prescribed or what complications might be expected in this half of the population. For example, the FDA first approved the drug zolpidem for the treatment of insomnia in the 1990s (brand name Ambien) and approved the use of generic versions in 2007. During both approval processes, scientific studies were carefully reviewed. But in 2013 the FDA issued the following statement:[19] “Since women eliminate zolpidem from their bodies more slowly than men, the FDA has notified the manufacturers that the recommended dose should be lowered for women.” For more than 20 years, women treated for insomnia were prescribed excessive dosages that often left them impaired the next day.

The National Institutes of Health (NIH) has also pointed out "an over-reliance on male animals, and neglect of attention to the sex of cells" in preclinical research using animal or cellular models. Starting in 2016, the NIH is requiring that all grant applications balance the sex of animals and cells (which carry either two X chromosomes or an X and Y chromosome set) used for research purposes and address sex as a biological variable. This move is part of the agency's ongoing effort to foster greater scientific rigor and generalizability in biomedical research.[20]

Statistical analysis of an experiment cannot tell us how far the results will generalize. The ability to extend conclusions beyond the experimental setting depends on a thorough discussion of the biological merits of the experiment and the representativeness of the subjects used. Nevertheless, a randomized comparative experiment, because of its ability to give convincing evidence for causation, is one of the most important ideas in statistics.

APPLY YOUR KNOWLEDGE

7.10 Reducing fat mass A 2015 peer-reviewed research article summarized its findings with both a bullet-point list of highlights and a graphical abstract. One bullet point reads, "Leptin stimulates lipolysis via sympathetic neurons in fat." The graphic above it shows the silhouettes of a clearly overweight woman and a slim woman, with the words "leptin" and "lipolysis" connecting the brain to fat deposits in the lower back, and an arrow implying change from the overweight silhouette to the slim silhouette.[21]

a. Go to http://dx.doi.org/10.1016/j.cell.2015.08.055 and read the freely available abstract. Can you tell from this research summary who or what the individuals used in the research study were?

b. The extensive research described in the article was entirely performed on mice, a fact that is clearly mentioned in the methods and results sections of the full article. Discuss the validity of summarizing the research findings with a graphical abstract representing a woman.

c. While the research article is very thorough in many respects, the sex of mice used in the research is not stated anywhere in the article. Explain why this information is relevant and should have been included.

7.11 Benefits of dark chocolate. Dark chocolate can dilate arteries by down-regulating the oxidative stress. A study investigated whether dark chocolate improves walking autonomy in patients with peripheral artery disease (PAD). The study compared the maximum walking distance (in meters) for 20 PAD patients after they ate 40 g of dark chocolate and also after they ate 40 g of milk chocolate, on two different weeks.[22] The published findings mention that “an individual not involved in the study randomly allocated the participants to a treatment sequence starting with dark chocolate or with milk chocolate, and kept this information in a sealed envelope. The authors and laboratory technicians were unaware of the treatment allocation.”

a. What are the individuals, the treatments, and the response variable in this experiment?

b. What is the design of this experiment? What type of blinding, if any, was used? Explain your answer.

STATS IN YOUR WORLD

SCRATCH MY FURRY EARS

Rats and rabbits, specially bred to be uniform in their inherited characteristics, are the subjects in many experiments. Animals, like people, are quite sensitive to how they are treated. This can create opportunities for hidden bias. For example, human affection can change the cholesterol level of rabbits. Choose some rabbits at random and regularly remove them from their cages to have their heads scratched by friendly people. Leave other rabbits unloved. All the rabbits eat the same diet, but the rabbits that receive affection have lower cholesterol.

ETHICS IN EXPERIMENTATION

We have dedicated the bulk of this chapter to describing what kinds of experiments *can* be done and discussing their respective scientific merits. We will now ask what kinds of experiments *should* or *should not* be done. This critically important aspect of experimental design is the domain of **ethics.**

Ethical issues are particularly important in biological experimentation because investigators deal with live forms. Whether the experimental units are humans, animals, or ecosystems, experimentation needs to be thought out not only to be statistically powerful and elegant but also to be *responsible.*

Experimental ethics is about all the questions we need to ask before, during, and after experimentation. But, unlike design, ethics doesn't always have clear-cut answers. Therefore, our objective here won't be to give you all the right answers but rather to help you ask all the right questions.

Human experimentation undoubtedly lends itself to particular ethical scrutiny. All experiments involving humans, even the least intrusive ones, raise ethical issues. Many experiments involving babies and young children have, for instance, helped us understand the nature and importance of attachment between a parent and a child. Yet we might legitimately ask whether the subjects themselves benefited from these experiments and whether they suffered any negative impacts. These same questions become paramount in the area of medical experimentation, where benefits and risks are clear and potentially great. Here are some questions we should ask ourselves:

- *Are all experiments acceptable in the name of science, knowledge, or the public's greater good?* Both medical ethics and international human rights standards say that "the interests of the subject must always prevail over the interests of science and society." The quoted words are from the 1964 Helsinki Declaration of the World Medical Association, the most respected international standard. The most outrageous examples of unethical experiments are those that ignore the interests of the subjects. Chief among them is the Tuskegee syphilis study, which is discussed later in this chapter.

- *Is it morally right to include a placebo in a study design when safe and efficient treatments are already available?* Because “the interests of the subject must always prevail,” medical treatments can be tested in clinical trials only when there is reason to hope that they will help the patients who are subjects in the trials. Future benefits aren’t enough to justify experiments with human subjects. So why is it ethical to give a control group of patients a placebo? Well, we know that placebos often work. What is more, placebos have no harmful side effects other than withholding the benefits of a proven treatment. If we knew which treatment was better, we would give it to everyone. When we don’t know, it is ethical to try both and compare them.
- *Should we stop an experiment if we find that one of the treatments shows early signs of adverse effects? And what if one of the treatments shows early signs of clear superiority?* Again, “the interests of the subject must always prevail.” Disregard of information gained in the course of experimentation or follow-up studies has been the basis of many private and class-action lawsuits against pharmaceutical companies and is fueling the current debate over the prescription of Vioxx, for instance.
- *Is it ever acceptable to leave subjects in the dark about the purpose and expected results of an experiment?* This question goes hand in hand with the first question on what experiments are morally acceptable. Subjects should be able to judge for themselves the acceptability of their participation. That’s why all federally funded human experiments are now required to obtain the subject’s *informed consent* in writing as well as the formal approval of an ethics committee. But can subjects truly grasp the purpose and consequences of an experiment? And is complete informed consent achievable without compromising the experiment? Some experiments, especially in psychology, must disguise their true purpose to obtain genuine responses from the participants.

EXAMPLE 7.15 **Studying drinking behavior**

> A study wanted to examine the impact of the shape of the glass (straight or curved) on how fast adult social drinkers finish a beer or a soft drink. Informing participants of this objective would likely change their natural drinking behavior, so the researchers obtained special approval from their university's Faculty of Science Research Ethics Committee to disguise the purpose of their study. The researchers told participants that the study examined "the effects of alcohol consumption on wordsearch performance," and they implemented the deception by having the participants actually complete a wordsearch after they had finished their drinks.[23]

- *Should subjects be rewarded for participation?* This is a tricky question related to informed consent and the freedom to choose participation. Poor and uninsured individuals could easily become primary targets for the riskiest experiments. In fact, it was once common to test new vaccines on prison inmates who gave their consent in return for good-behavior credit. The law now forbids medical experiments in prisons. And, of course, we must also learn to think globally because new treatments can be tested anywhere, including in parts of the world where subjects' rights have not yet fully matured.

By law, all studies *funded by the federal government* must obey the following principles, which are collectively referred to as the "Common Rule":

- The organization that carries out the study must have an *institutional review board* that reviews all planned studies in advance to protect the subjects from possible harm.
- All individuals who are subjects in a study must give their *informed consent* in writing before data are collected. Subjects must be informed about the nature of a study and any risk of harm it may bring.
- All individual data must be kept *confidential*. Only statistical summaries for groups of subjects may be made public. Any breach of confidentiality is a serious violation of data ethics. The best practice is to separate the identity of the subjects from the rest of the data at once.

EXAMPLE 7.16 Who can consent?

Children can’t give fully informed consent, so the usual procedure is to ask their parents. A study of new ways to teach reading is about to start at a local elementary school, so the study team sends consent forms home to parents. Many parents don’t return the forms. Can their children take part in the study because the parents did not say “no,” or should we obtain data from only those children whose parents returned the form and said “yes”?

What about research into new medical treatments for people with mental disorders? What about studies of new ways to help emergency room patients who may be unconscious? In most of the latter cases, there isn’t time to get the consent of the family. Does the principle of informed consent bar realistic trials of new treatments for unconscious patients?

These are questions without clear answers. Reasonable people differ strongly on their opinions on all of them. Not everything about informed consent is simple.[24]

While it is important to question the necessity and ethicality of human and animal experimentation, we also must consider the implications of not performing carefully designed experiments. There are many unfortunate examples of pharmaceutical or surgical methods that have been or still are widely used without going through this important preliminary stage. In such a situation, the whole population becomes unsuspecting test subjects of an “experiment” that has no guidelines, no safeguards, and no way to reach informative conclusions about the benefits and dangers of the drug or procedure.

Lobotomy is the severance of the frontal lobes from the rest of the brain. In the 1950s and 1960s, it became a popular method for dealing with psychotic individuals. Lobotomy was performed on consenting, somewhat-informed individuals or with the approval of family members wishing to stop undesirable behavior (even in children). After thousands of lobotomies were performed in the United States alone, it became clear that the procedure did provide improvement for some patients, but others showed no improvement at all, and the rest suffered irreversible, lifelong psychological damage.[25]

This is not an isolated case. Surgical practice, especially, has evolved largely based on the ideas and attempts of individual surgeons. Because surgery is by definition invasive, there has historically been a reluctance to perform randomized, controlled experiments in this field out of concern for the subjects. However, with any new procedure, a placebo or sham operation may prove equally or more beneficial than the procedure tested and have many fewer or milder side effects. Gastric freezing for the treatment of stomach ulcers (described in Example 7.3) and mammary artery ligation for the treatment of angina (described in Exercise 15.12, page 390) are two medical procedures that were widely used before randomized comparative experiments showed that their results were not any better than those obtained with placebos, leading to the procedures being abandoned. Likewise, the use of routine episiotomy to speed up labor and prevent natural tearing of the perineum during childbirth has been questioned by multiple studies,[26] and its use has been declining rapidly. Fortunately, these drastic practice reversals have led to a greater awareness of the necessity for careful experimentation to legitimize the use of medical and surgical procedures in humans.

DISCUSSION The Tuskegee syphilis study

F“or 40 years, the U.S. Public Health Service has conducted a study in which human guinea pigs, not given proper treatment, have died of syphilis and its side effects. The study was conducted to determine from autopsies what the disease does to the human body.” The *Washington Evening Star* published these lines in a front-page article in the summer of 1972. The article would end what was perhaps the most controversial human experiment ever in the United States and result in drastic changes to federal regulations related to scientific and medical experimentation requiring human subjects.

The Public Health Service and the Tuskegee Institute began a study of syphilis in 1932. Nearly 400 poor black men with syphilis from Macon County, Alabama, were enrolled in the study. They were never told they had syphilis and were never treated for it—even after penicillin became known as a highly effective cure by 1943 and widely available by 1947. Also, they were never told that they were part of a medical experiment.

Syphilis

Syphilis is a sexually transmitted disease caused by the spirochete bacterium *Treponema pallidum*. Its pathology progresses in stages, starting with the development of hard ulcers primarily around the genitalia, which is followed by a transient, widespread rash. If the person does not receive treatment, the third stage of syphilis may develop, damaging the nervous system, eyes, heart, blood vessels, liver, bones, and joints, causing blindness, dementia, and paralysis.

Untreated men can transmit the disease to their female partners, who can then transmit it to their future offspring. Approximately 40% of pregnancies where the mother has untreated syphilis result in stillbirth (baby dead at birth). If the baby survives birth, it then has a 40% to 70% chance of being infected with syphilis. If left untreated, infected newborns may later become developmentally disabled, have seizures, or die because of the disease.

How could this happen?

How could a study designed to deny treatment to uninformed subjects have occurred? The historical context can help us understand the logical journey.

Courtesy of the national Archives at Atlanta

At the end of the 19th century, philanthropists dreamed of black economic development, and the Tuskegee Institute was founded to help develop local schools, businesses, and agriculture. Its mission was later expanded to fight the devastating impact of a syphilis pandemic that affected nearly one-third of the African American population in the South. At that time, diagnosis and the currently known treatment were offered to participants.

When the stock market crashed in 1929, funds for the endeavor were cut. In 1932, the Tuskegee philanthropic effort was recycled into a study of the impact of untreated syphilis. In exchange for participating in the study, the men were given free medical exams, free meals, and free burial insurance. The study was originally intended to run for six months, but it lasted 40 years, until a whistleblower turned to the press.

Acknowledging misconduct

The Tuskegee study was well known in the scientific community, and several papers on its findings were published in peer-reviewed journals. The first concerns over ethical aspects of the study were raised in 1968, but the study was terminated only after the outrage generated by the 1972 press release. It then took 25 more years for the government to voice a formal apology. Here is an excerpt from President Bill Clinton’s 1997 formal apology on behalf of the nation (the full transcript can be read at http://clinton4.nara.gov/textonly/New/Remarks/Fri/19970516-898.html):

> *The United States government did something that was wrong—deeply, profoundly, morally wrong. It was an outrage to our commitment to integrity and equality for all our citizens.*
>
> *To the survivors, to the wives and family members, the children and the grandchildren, I say what you know: No power on Earth can give you back the lives lost, the pain suffered, the years of internal torment and anguish. What was done cannot be undone. But we can end the silence. We can stop turning our heads away. We can look at you in the eye and finally say on behalf of the American people, what the United States government did was shameful, and I am sorry.*

How the Tuskegee study changed the ethics of human experimentation

The revelation of the Tuskegee syphilis study was the spark that initiated profound and long-lasting reforms of the ethical code of conduct for research involving human subjects in the United States. In 1974, the National Research Act was signed into law, creating the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Today, all federally supported studies using human subjects must be reviewed by institutional review boards, which decide whether study protocols meet ethical standards, and researchers must get voluntary informed consent from all study participants.

Complementing this approach, training for researchers now emphasizes being aware of the ethical issues related to human experimentation. For instance, the National Institutes of Health now requires that all prebaccalaureate, predoctoral, and postdoctoral trainees supported by institutional training grants receive instruction in the responsible conduct of research.

Despite the ethical progress made in regard to the conduct of experiments in research institutions, experiments performed by private entities are not actually bound by those principles. Exercise 7.13 cites an example of legally acceptable, yet ethically questionable experimentation that was privately funded by a corporation.

Data privacy is another concerning issue with inconsistent ethical practices. Wearables and apps are increasingly common tools used to monitor personal health data (think step counters, sleep monitoring programs, calorie calculators, and medication schedulers). Unless they are programmed to send your data directly to your official health care provider, such health tools are not covered by the Health Information Portability and Accountability Act (HIPAA).[27] Any promise of confidentiality for your data, then, is entirely voluntary. Even basic online searches and in-store purchase data can be used to deduce very private medical information. In a now infamous case, the giant retailer Target once deduced that a teenager was pregnant before anyone else in her family had been informed, and the very specific coupons she received in the mail created a fair amount of family mayhem and national news coverage.[28]

Experimentation on animals, ecosystems, cells, and embryos also raises many ethics questions, though standards in these cases are much more variable. Federal agencies also regulate the use of animals for laboratory experiments, but the guidelines and scrutiny applied are very different for different species. For example, primate experiments in the United States are being increasingly more tightly regulated, in drastic contrast to research on insects. Rats are the typical animal model used for testing suspected carcinogenicity of various substances. Such experiments obviously do not have the animals' well-being in mind. Rats, however, are considered pests and are routinely exterminated by farmers and homeowners, and we consider the information gained from carcinogenicity studies to be important for the well-being and safety of entire human populations. Rat experiments used to test beauty products, in contrast, have been much more controversial. Animal-rights activists advocate giving all animals (or at least species with some level of self-consciousness) the same ethical consideration as humans. The debate over animal experimentation is still far from being settled, and this state of flux is exemplified by the various laws, regulations, and practices enforced in different countries and different points in time.

Even the status of cells has been a subject of ethical controversy when these cells have the potential to become human beings. Stem cell research and its applications have been debated all over the world, with varying results. Within the United States, the debate has changed over time. In 2001, President George W. Bush changed the U.S. policy on human embryonic stem cell research by allowing federal funds to be used for research only on existing stem cell lines.[29] In 2004, California passed a stem cell initiative allowing for $3 billion in funding over 10 years for stem cell research and research facilities in that state.[30] In 2009, President Barack Obama lifted the 2001 restrictions on stem cell research.[31] And in 2016, the NIH proposed lifting an earlier ban on funding experiments that use human stem cells to create human–animal embryos for carefully monitored conditions.[32]

The answers to questions of ethics in experimentation are rarely easy. But perhaps as important as the decisions we make is keeping the debate alive and open to public discussion.

An alternative to actual experimentation that has tremendous potential is to run “experiments” on the computer instead. As computers become ever more powerful and with our knowledge of the world ever deepening, we can *simulate* everything we know about a system and the forces that regulate it. The system can then be modified and observed, just as we would apply a treatment and observe the outcome in the real world.

APPLY YOUR KNOWLEDGE

7.12 Should we experiment on newborns? The United States is in the midst of a large and growing opioid epidemic, with individuals’ addictions often being initiated by an opioid prescription intended to treat a medical condition. As a result, the number of newborns experiencing the intense withdrawal symptoms of neonatal abstinence syndrome have greatly increased in recent years. What medical care should be provided to these newborns? Here is a portion of a 2016 NPR news article that discusses the issues:[33]

> *“The American Academy of Pediatrics and others recommend an opioid for the babies, because you’re giving them back what they’re withdrawing from,” explains Dr. Jonathan Davis, a neonatologist and chief of newborn medicine at Tufts’ Medical Center. “Morphine and methadone are the two most common.”*

> *But Davis says no one's really done the research to figure out which drug works better for babies, and doctors are left to figure that out by trial and error, case by case. Though the Food and Drug Administration hasn't officially approved morphine or methadone for use in newborns, doctors prescribe these drugs to the children anyway, in smaller doses than they give adults.*
>
> *"As I spoke to people around the country, everyone would have their own approach and a very different way of treating these babies," Davis says. "And we thought that quite odd."*
>
> *So he and a colleague, Brown University developmental psychologist Barry Lester, have launched a major study to sort out what works best. The two are hoping to enroll 180 babies in their double-blind, randomized, controlled trial—no one will know which newborns are getting methadone, and which are getting morphine, for example, until the study's end.*

a. Describe some of the ethical concerns of experimenting on newborns in the context of neonatal abstinence syndrome.

b. Explain why the proposed experiment may nonetheless be ethically acceptable.

c. An online commenter hoped that no one would be given a placebo. Discuss why a placebo would or would not be ethically appropriate in the context of neonatal abstinence syndrome.

7.13 When Facebook experiments. We know that emotional states can be transferred to others via emotional contagion. But does emotional contagion occur outside of in-person interactions? A research article describes the findings of an experiment conducted by Facebook for internal purposes. The experiment manipulated the amount of emotional content in the posts from friends shown in the news feeds of 689,003 Facebook users. Based on random assignment, these Facebook users received news feeds that were either modified to have reduced positive expression, modified to have reduced negative expressions, or left intact. When positive expressions were reduced, Facebook users posted fewer positive words and more negative words; when negative expressions were reduced, the opposite pattern occurred.[34]

a. How were participants selected for this experiment? Here is what the article says: “People who viewed Facebook in English were qualified for selection into the experiment.” No notification was provided for those 689,003 qualified participants, and they were not given any opportunity to opt out of the study. Describe some ethical concerns about such an approach to human experimentation.
b. The article states that participation in the experiment “was consistent with Facebook’s Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research.” Explain how agreeing to Facebook’s terms and conditions at the time of account creation differs from the ethics requirements for human research referred to as the “Common Rule.”
c. The effect resulting from the emotional manipulation was very small, on the order of 1 fewer emotional word per 1000 words posted the following week. Does this fact affect the validity of the ethical concerns over this experiment?
d. Many scientists voiced their concern in the media over the ethicality of this experiment, and the editor-in-chief of the scientific journal in which this article was published demanded that an editorial expression of concern be added to the article. How is this involvement from the scientific community different from the scientific community’s reaction to publications about the Tuskegee syphilis study decades earlier (described in the Discussion on page 185).

CHAPTER 7 SUMMARY

- In an experiment, we impose one or more **treatments** on individuals, often called **subjects** or **experimental units.** Each treatment is a combination of values of the explanatory variables, which we call **factors.**
- The **design** of an experiment describes the choice of treatments and the manner in which the subjects are assigned to the treatments.
- The basic principles of statistical design of experiments are **control** and **randomization** to combat bias and **using enough subjects** to reduce chance variation.
- The simplest form of control is **comparison.** Experiments should compare two or more treatments to avoid **confounding** of the effect of a treatment with other influences, such as lurking variables.
- **Randomization** uses chance to assign subjects to the treatments. Randomization creates treatment groups that are similar (except for chance variation) before the treatments are applied. Randomization and comparison together prevent **bias,** or systematic favoritism, in experiments. Randomization also allows the use of statistical analyses based on laws of probability.
- You can carry out randomization by using software or by giving numerical labels to the subjects and using a table of random digits to choose treatment groups.
- Applying each treatment to many subjects reduces the role of chance variation and makes the experiment more sensitive to differences among the treatments.
- Good experiments require attention to detail as well as good statistical design. Many behavioral and medical experiments are **double-blind.** Some give a **placebo** to a control group or introduce a **double-dummy** when the treatments compared are too visibly different for a simple placebo. **Replicating** an experiment helps ensure that the results were not merely due to unusual conditions. **Lack of realism** in an experiment

can prevent us from generalizing its results.

- In addition to comparison, a second form of control is to restrict randomization by forming **blocks** of individuals that are similar in some way that is relevant to the response. Randomization is then carried out separately within each block.
- **Matched pairs** are a common form of blocking for comparing just two treatments. In some matched pairs designs, the subjects are matched in pairs as closely as possible, and each subject in a pair receives one of the treatments. In other designs, each subject receives both treatments in a random order, an experimental design also called **repeated measures.**
- In the end, **ethics** must be the overarching principle guiding experimentation. No matter which experiments could be done and what we might gain from them, we need to ask ourselves if the end justifies the means and be open to a continuing debate.

THIS CHAPTER IN CONTEXT

Observational studies and experiments are two methods for producing data. Observational studies are useful when the conclusion involves describing a group or situation without disturbing the scene we observe. In contrast, experiments are used when the situation calls for a conclusion about whether a treatment *causes* a change in a response. The distinction between observational studies and experiments will be important when stating your conclusions in context in all chapters on statistical inference (Chapters 14 through 28).

Only well-designed experiments provide a sound basis for concluding that a cause-and-effect relationship exists. In the simplest comparative experiment, two treatments are imposed on two groups of individuals. Reaching the conclusion that the resulting difference between the groups is caused by the treatments rather than by lurking variables requires that the two groups of individuals be similar at the outset. A randomized comparative experiment is used to create groups that are similar. If a sufficiently large difference between the groups is found after imposing the treatments, we can say that the results are statistically significant and conclude that the differences in the response were most likely *caused* by the treatments. In Parts III and IV, the specific statistical procedures for reaching these conclusions will be described. We will also see that which statistical procedure should be used depends in part on the type of experimental design used to collect the data.

As with sampling, when conducting an experiment, attention to detail is important because our conclusions can be weakened by several issues. A lack of blinding can result in the researcher’s expectations influencing the results, while the placebo effect can confound the comparison between a treatment group and a control group. In many instances, a more complex design is required to overcome difficulties and can produce more precise results. Such issues should be addressed before starting data collection and may warrant assistance from a professional statistician.

CHECK YOUR SKILLS

7.14 What electrical changes occur in muscles as they get tired? Student subjects hold their arms above their shoulders until they drop. Meanwhile, the electrical activity in their arm muscles is measured. This is

a. an observational study.
b. an uncontrolled experiment.
c. a randomized comparative experiment.

7.15 Can changing a person’s diet reduce high blood pressure? Vegetarian diets and low-salt diets are both promising alternatives. Men with high blood pressure are assigned at random to four diets: (1) normal diet with unrestricted salt, (2) vegetarian diet with unrestricted salt, (3) normal diet with restricted salt, and (4) vegetarian diet with restricted salt. This experiment has

a. one factor—the choice of diet.
b. two factors—normal/vegetarian diet and unrestricted/restricted salt.
c. four factors—the four diets being compared.

7.16 The response variable in the experiment described in Exercise 7.15 is

a. the amount of salt in the subject’s diet.
b. which of the four diets a subject is assigned to.
c. change in blood pressure after 8 weeks on the assigned diet.

7.17 In the experiment described in Exercise 7.15, the 240 subjects are labeled 001 to 240. Software assigns an SRS of 60 subjects to Diet 1, then an SRS of 60 of the remaining 180 subjects to Diet 2, then an SRS of 60 of the remaining 120 subjects to Diet 3. The 60 subjects who are left get Diet 4. This is a

a. completely randomized design.
b. block design, with four blocks.
c. matched pairs design.

7.18 A study of the relationship between dengue infection and miscarriage enrolled a group of women who had just experienced a miscarriage. The investigators matched each woman who had suffered a miscarriage to three women with viable pregnancies at a similar gestational stage, then asked them if they had had a recent dengue infection.[35] This is

a. an observational study with a case-control design.
b. an experiment with a block design.
c. an experiment with a matched pairs design.

7.19 A medical experiment compares an antidepressant with a placebo for relief of chronic headaches. A total of 38 patients with chronic headaches are available to serve as subjects: 28 women and 10 men. The investigators randomly select 16 female patients to receive the antidepressant and they also randomly select 5 men to receive the antidepressant. The remaining patients will receive the placebo. This experiment has

a. a completely randomized design.
b. a matched pairs design.
c. a block design.

7.20 Can goats remember and identify the call of their kid (baby goat)? Researchers selected 9 mother goats who had weaned their kids nearly a year earlier. They measured the intensity of the mother's response when presented with the recorded call of their own kid and also with the recorded call of a different kid. They found that the mothers' responses were significantly more intense for their own kid.[36] This is

a. an observational study.
b. a matched pairs experiment.
c. a completely randomized experiment.

7.21 To decide which call the mother goats in Exercise 7.20 should hear first, it is best to

a. always start with the same call (for example, the different kid) for consistency.
b. base the decision on how attached each mother goat was to her kid.
c. use chance (for example, a coin toss).

7.22 The National Institutes of Health should be reluctant to use the results of the experiment described in Exercise 7.15 for its guideline on diets for Americans with high blood pressure because

a. the study used a matched pairs design instead of a completely randomized design.
b. results from men with high blood pressure may not generalize to the population of all Americans with high blood pressure.
c. the study did not use a placebo and was not double-blind.

7.23 An experimental design randomly assigns volunteer patients suffering from chronic pain due to arthritis to take either a new pain medication, aspirin (the traditional nonsteroidal anti-inflammatory treatment), or a placebo. An institutional review board discusses whether the experiment should include a placebo. This is

a. an ethical issue.
b. an issue of lack of replication.
c. an issue of lack of realism.

CHAPTER 7 EXERCISES

In exercises that require randomization, you may use any appropriate software, the Simple Random Sample applet, or Table A.

7.24 Dental bonding. A dentistry study evaluated the effect of tooth etch time on resin bonding strength. A total of 78 undamaged, recently extracted first molars (baby teeth) were randomly assigned to be etched with phosphoric acid gel for either 15, 30, or 60 seconds. Composite resin cylinders of identical size were then bonded to the tooth enamel. The researchers examined the bond strength after 24 hours by finding the failure load (in megapascals) for each bond.

a. What are the explanatory and response variables?
b. Outline the design of this experiment.

7.25 Treating breast cancer. The most common treatment for breast cancer discovered in its early stages was once removal of the breast. Today, the usual practice is to remove only the tumor and nearby lymph nodes, with the surgery then being followed by radiation. To study whether these treatments differ in their effectiveness, a medical team examines the records of 25 large hospitals and compares the survival times after surgery of all women who have had either treatment.

a. What are the explanatory and response variables?
b. Explain carefully why this study is not an experiment.
c. Explain why confounding will prevent this study from discovering which treatment is more effective. (The current treatment was recommended after several large randomized comparative experiments.)

7.26 Growing in the shade. Ability to grow in shade may help pines in the dry forests of Arizona resist drought. How well do these pines grow in shade? Investigators planted 45 pine seedlings in a greenhouse in either full light, light reduced to 25% of normal by shade cloth, or light reduced to 5% of normal. At the end of the study, they dried the young trees and weighed them.

a. What are the individuals, the treatments, and the response variable in this experiment?
b. Outline the design of this experiment. Use software or Table A to randomly assign seedlings to the treatment groups.

7.27 Exercise and heart attacks. Does regular exercise reduce the risk of a heart attack? Here are two ways to study this question. Explain clearly why the second design will produce more trustworthy data.

1. A researcher finds 2000 men older than age 40 who exercise regularly and have not had heart attacks. She matches each with a similar man who does not exercise regularly, and she follows both groups for five years.
2. Another researcher finds 4000 men older than age 40 who have not had heart attacks and are willing to participate in a study. She assigns 2000 of the men to a regular program of supervised exercise. The other 2000 continue their usual habits. The researcher follows both groups for five years.

7.28 Treating brain cancer, continued. Researchers had 22 mice available for the experiment on brain cancer treatment described in Exercise 7.2. They assigned 4 to the irradiation condition, 5 to the cannabis condition, 6 to the irradiation combined with cannabis, and the remaining 7 to the no treatment condition. Use a diagram to outline a completely randomized design for this experiment. Do the randomization required to assign mice to the conditions using software, the *Simple Random Sample* applet or Table A (start at line 105).

APPLET

7.29 Bird songs. Bird songs have been hypothesized to be a secondary sexual character signaling an individual’s health status. Researchers designed an experiment in which they randomly assigned male collared flycatchers (*Ficedula albicollis*) to two groups. One group received an immune challenge in the form of an injection of sheep red blood cells, and the other group received a placebo injection. The researchers then recorded the changes in song rate (in strophes per minute)after the injection for the

15 males in the immune challenge group and the 12 males in the placebo group.

a. Outline the design of this experiment.
b. Label the birds from 1 through 27. Use software or Table A at line 131 to carry out the randomization your design requires.

7.30 Beehives. Researchers created 14 artificial see-through beehives (using transparent Plexiglas) that can be heated on demand with infrared lamps. The researchers videotaped the bees, both before and after heating the beehive, to count how many bees in the hive engage in heat-reduction behavior (“heat shielders”) under each condition.[37]

a. Identify the individuals, the factors, the response variable, and the design used for this study.
b. Follow Figure 7.6 as a guideline to construct an appropriate table to store the data from this experiment.
c. The published report states that “there were significantly more heat shielders observed after heating than before heating (p=0.01).$(p = 0.01)$.” Explain what “[statistically] significantly” means in the context of this study.

7.31 Relieving headaches. Doctors identify “chronic tension-type headaches” as headaches that occur almost daily for at least six months. Can antidepressant medications or stress management training reduce the number and severity of these headaches? Are both together more effective than either alone?

a. Create a diagram like Figure 7.1 to display the treatments in a design with two factors: medication (yes or no) and stress management (yes or no). Then outline the design of a completely randomized experiment to compare these treatments.
b. The headache sufferers named here have agreed to participate in the study. Randomly assign the subjects to the treatments. If you use the *Simple Random Sample* applet or other software, assign all the subjects. If you use Table A, start at line 130 and assign subjects to only the first treatment group.

APPLET

Abbott
Abdalla
Alawi
Broden
Chai
Chuang
Cordoba
Custer
Decker
Devlin
Engel
Fuentes
Garrett
Gill
Glover
Hammond
Herrera
Hersch
Hurwitz
Irwin
Jiang
Kelley
Kim
Landers
Lucero
Masters
Morgan
Nelson
Nho
Ortiz
Ramdas
Reed
Richter
Riley
Samuels
Smith
Suarez
Upasani
Wilson
Xiang

7.32 Oligofructose and calcium absorption. Nondigestible oligosaccharides are known to stimulate calcium absorption in rats. A double-blind, randomized experiment investigated whether the consumption of oligofructose similarly stimulates calcium absorption in 11 healthy male adolescents 14 to 16 years old. The subjects took a pill for nine days and had their calcium absorption tested on the last day. The experiment was repeated three weeks later. Some subjects received the oligofructose pill in the first round and then a pill containing sucrose (serving as a control) in the second round. The order was switched for the remaining subjects.

a. What are the explanatory and response variables here?
b. Which design was used in this experiment? Why was the order of treatment switched for some subjects?
c. The study was double-blind. What does this mean?

7.33 Cell phone use and brain metabolism. A study examined the impact of cell phone use on brain metabolism in 47 healthy adults. Here is a quote from the published abstract: “Cell phones were placed on the left and right ears and positron emission tomography with (^{18}F)fluorodeoxyglucose injection was used to measure brain glucose metabolism twice, once with the right cell phone activated (sound muted) for 50 minutes (ON condition) and once with both cell phones deactivated (OFF condition).”[38]

a. This experiment has a matched pairs design. How can you tell?
b. What is the advantage of using a matched pairs design rather than a completely randomized design in this context?

c. The abstract reports that “metabolism in the region closest to the antenna (orbitofrontal cortex and temporal pole) was significantly higher for ON than OFF conditions.” Explain in simple terms what “significantly higher” means in this context.

7.34 Athletes taking oxygen. We often see players on the sidelines of a football game inhaling oxygen. Their coaches think this treatment will speed their recovery. We might measure recovery from intense exercise as follows. Have a football player run 100 yards three times in quick succession. Then allow three minutes to rest before running 100 yards again. Time the final run. Because players vary greatly in speed, you plan a matched pairs experiment using 25 football players as subjects.

a. Discuss the design of such an experiment to investigate the effect of inhaling oxygen during the rest period. What is the name for this experimental design?
b. Use a diagram to outline the design.

7.35 Protecting ultramarathon runners. An ultramarathon, as you might guess, is a footrace longer than the 26.2 miles of a marathon. Runners commonly develop respiratory infections after an ultramarathon. Will taking 600 milligrams of vitamin C daily reduce these infections? Researchers randomly assigned ultramarathon runners to receive either vitamin C or a placebo. Separately, they also randomly assigned these treatments to a group of nonrunners the same age as the runners. All subjects were watched for 14 days after the big race to see if infections developed.

a. What is the name for this experimental design?
b. Use a diagram to outline the design.

7.36 Reducing spinal fractures. Fractures of the spine are common and serious among women with advanced osteoporosis (low mineral density in the bones). Can taking strontium renelate help? A large medical trial assigned 1649 women to take either strontium renelate or a placebo each day. All the subjects had osteoporosis and had suffered at least one fracture. All were taking calcium supplements and receiving standard medical care. The response variables were measurements of bone density and counts of new fractures over three years. The subjects were treated at 10 medical centers, each in a different country. Outline a block design for this experiment, with the medical centers as blocks. Explain why this is the proper design.

7.37 Does ginkgo improve memory? The law allows marketers of herbs and other natural substances to make health claims that are not supported by evidence. Brands of ginkgo extract claim to "improve memory and concentration." A randomized comparative experiment found no evidence for such effects. The subjects were 230 healthy people older than age 60 years. They were randomly assigned to receive ginkgo or a placebo pill. All the subjects took a battery of tests for learning and memory before treatment started and again after six weeks.

a. Following the model of Figure 7.3, outline the design of this experiment.

b. Use the *Simple Random Sample* applet, other software, or Table A to assign half the subjects to the ginkgo group. If you use software, report the first 20 members of the ginkgo group (in the applet's "Sample bin") and the first 20 members of the placebo group (those left in the "Population hopper"). If you use Table A, start at line 103 and choose only the first 5 members of the ginkgo group.

7.38 Quick randomizing? Here's a quick and easy way to randomize. You have 100 subjects, 50 women and 50 men. Toss a coin. If it's heads, assign the men to the treatment and the women to the control group. If the coin comes up tails, assign the women to treatment and the men to control. This gives every individual subject a 50-50 chance of being assigned to treatment or control. Why isn't this a good way to randomly assign subjects to treatment groups?

7.39 "Doctor offers women flashy new treatment." Here is an excerpt from a 2007 NBC news report:

> *Millions of women suffer from severe mood swings and hot flashes associated with menopause and other causes of estrogen depletion. . . . Lipov is treating menopausal women by using an injection which he believes can reset the brain's thermometer to reduce or eliminate hot*

> *flashes. . . . Lipov is the first and only doctor in the U.S. to target hot flashes. To make the procedure more safe, he first injects a dye into the targeted area. Complications, like seizures, are rare and so far Lipov's patients have had no problems. The hot flash relief lasts from weeks to months, Lipov said, and almost all of the 22 women he's used the technique on have gotten some benefit.*[39]

a. What is the design of this study?
b. Can we trust the results of this study? Explain why or why not.

7.40 Antimicrobial agent in consumer products. Triclosan (TCS) is a broad-spectrum antimicrobial agent commonly added to soaps and other consumer products. TCS is so ubiquitous that it has frequently been detected in human body fluids. Researchers examined the long-term effect of exposure to TCS in an animal model. They randomly assigned three-week-old male mice to diets containing either 0.08% TCS or no TCS for eight months, before performing a number of histological and immunochemical analyses on the mice.[40]

a. Identify the individuals, the factors, and the broad response variables in this study. What design was used?
b. Here is a quote from the published report: "Although animal studies require higher chemical concentrations than predicted for human exposure, this study demonstrates that TCS acts as a HCC tumor promoter and that the mechanism of TCS-induced mouse liver pathology may be relevant to humans." What lack of realism aspect of this study is discussed in this quote?
c. Identify one other way in which this study may not necessarily be "relevant to humans."

7.41 Explaining medical research. Observational studies had initially suggested that vitamin E may reduce the risk of heart disease. Careful experiments, however, showed that vitamin E has no effect, at least for women. According to a commentary in the *Journal of the American Medical Association,* vitamin E is yet another treatment that may have been "promising in epidemiologic and observational studies but failed to deliver in adequately powered randomized controlled trials." The author attributes this reversal to a healthy-user bias, “the healthy lifestyle behaviors that characterize individuals who care enough about their health to take various supplements [and] are actually responsible for the better health.”[41] A friend who knows no statistics asks you to explain this commentary.

a. What is the difference between observational studies and experiments?
b. What is a “randomized controlled trial”? (We’ll discuss “adequately powered” in Chapter 15.)
c. How does a healthy-user bias explain how people who take vitamin E supplements have better health in observational studies but not in controlled experiments?

7.42 Treating sinus infections. Sinus infections are common, and doctors often treat them with antibiotics. Another treatment is to spray a steroid solution into the nose. A well-designed clinical trial found that these treatments, alone or in combination, do not reduce the severity or the length of sinus infections. The clinical trial was a completely randomized experiment that assigned 240 patients at random among 4 treatments as follows:

	Antibiotic pill	**Placebo pill**
Steroid spray	53	64
Placebo spray	60	63

a. What are the factors and treatments in this experiment?
b. Outline the design of the experiment.
c. Explain briefly how you would do the random assignment of patients to treatments.

7.43 Treating sinus infections, continued. If the random assignment of patients to treatments in Exercise 7.42 did a good job of eliminating bias, possible lurking variables such as smoking history, asthma, and hay fever should be similar in all 4 groups. After recording and comparing many such variables, the investigators said that “all showed no significant difference between groups.” Explain to someone who knows no statistics what “no significant difference” means. Does it mean that the presence of all these variables was exactly the same in all four treatment groups?

7.44 Growing tomato plants. An experiment measures the amounts of various nutrients in tomato plants subjected to three levels of nitrogen fertilization. To avoid confounding by the soil properties, the planting area was divided into 12 blocks within which subplots were randomly assigned to the 3 treatments. Draw an outline of this block design in the style of Figure 7.5.

7.45 An herb for depression? Does the herb Saint-John’s-wort relieve major depression? A study examined this issue. Here are some excerpts from the published report:[42]

> *Design: Randomized, double-blind, placebo-controlled clinical trial. … Participants …were randomly assigned to receive either Saint-John’s-wort extract (n = 98) or placebo (n = 102). …The primary outcome measure was the rate of change in the Hamilton Rating Scale for Depression over the treatment period.*

a. Use the information provided to draw a diagram outlining the design of this study.
b. The report concluded that Saint-John’s-wort is no more effective than a placebo at relieving symptoms of depression. Based on the information quoted, would you be inclined to trust this conclusion? Explain your answer.

7.46 Treating Lyme disease. A study found that longer-term antibiotic treatment of persistent Lyme disease symptoms does not lead to better outcomes than shorter-term treatment. Here is a paragraph taken from the published report:[43]

> *We assigned patients with persistent symptoms attributed to Lyme disease to receive a 12-week oral course of doxycycline, clarithromycin plus hydroxychloroquine, or placebo. All study groups received open label intravenous ceftriaxone for 2 weeks before*

initiating the randomized regimen. The primary outcome measure was health-related quality of life, as assessed by the physical-component summary score of the RAND-36 Health Status Inventory (range, 15 to 61, with higher scores indicating better quality of life), at the end of the treatment period. All personnel involved in the study (except the data manager and pharmacist) and all participants were unaware of the study-group assignments.

a. Identify the individuals, the factors, and the response variable in this study.
b. Draw a diagram outlining the design of this study.
c. Identify all the steps taken by the researchers to avoid bias.
d. Explain why the researchers treated all subjects with intravenous ceftriaxone before randomization?

7.47 Experimentation for emergency treatments. How do you deliver oxygen to cells in emergency situations when the lungs have failed? A team of researchers devised a new method that encapsulates oxygen within microparticles of fat so that oxygen can be injected directly into the bloodstream without risking an embolism. The method was tested experimentally in rabbits and shown to be successful.[44] But how do you then test it in humans?

a. One step would involve giving the oxygen injections to healthy volunteers to test its safety. What are some ethical challenges to using healthy volunteers to test an unknown medical procedure?
b. Another step would require giving the oxygen injections during real-life emergencies involving oxygen deprivation. Obtaining informed consent from the patient is likely to be impossible, and there may not be enough time to obtain informed consent from family members. Would it be ethical to test the procedure anyway?

7.48 Institutional review boards. Your college or university has an institutional review board that screens all studies that use human subjects. Get a copy of the document that describes this board (you can probably find it online).

a. According to this document, what are the duties of the board?

b. How are members of the board chosen? How many members are not scientists? How many members are not employees of the college? Do these members have some special expertise, or are they simply members of the "general public"?

7.49 World ethics. Researchers from Yale, working with medical teams in Tanzania, wanted to know how common infection with the human immunodeficiency virus (HIV), which causes acquired immunodeficiency syndrome (AIDS), is among pregnant women in that African country. To do this, they planned to test blood samples drawn from pregnant women. Yale's institutional review board insisted that the researchers get the informed consent of each woman and tell her the results of the test. This is the usual procedure in developed nations. The Tanzanian government did not want to tell the women why blood was drawn or tell them the test results. The government feared panic if many people turned out to have an incurable disease for which the country's medical system could not provide care. The study was canceled. Do you think that Yale was right to apply its usual standards for protecting subjects?

7.50 What is minimal risk? Federal regulations say that "minimal risk" means the risks are no greater than "those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests." You area member of your college's institutional review board. You must decide whether several research proposals qualify for lighter review because they involve only minimal risk to subjects. Which of these do you think qualifies as "minimal risk"?

a. Draw a drop of blood by pricking a finger to measure blood sugar.
b. Draw blood from the arm for a full set of blood tests.
c. Insert a tube that remains in the arm, so blood can be drawn regularly.

7.51 The Willowbrook hepatitis studies. In the 1960s, children entering the Willowbrook State School, an institution for the developmentally disabled, were deliberately infected with hepatitis. The researchers argued that almost all children in the institution quickly became infected anyway. The studies showed for the first time that two strains of hepatitis existed. This finding contributed to the development of effective vaccines. Despite these valuable results, the Willowbrook studies are now considered an example of unethical research. Explain why, according to current ethical standards, useful results are not enough to allow a study.

7.52 Informed consent. Long ago, doctors drew a blood specimen from you as part of treating minor anemia. Unknown to you, the sample was stored. Now researchers plan to used stored samples from you and many other people to look for genetic factors that may influence anemia. It is no longer possible to ask your consent. Modern technology can read your entire genetic makeup from the blood sample.

a. Do you think it violates the principle of informed consent to use your blood sample if your name is on it but you were not told that it might be saved and studied later?
b. Suppose that your identity is not attached. The blood sample is known only to come from (say) “a 20-year-old white female being treated for anemia.” Is it now acceptable to use the sample for research?
c. In 2015, the NIH requested comments on the Common Rule for human experimentation. One proposed change is to “require consent for any use of biospecimens, regardless of whether personal identifiers are attached to the biospecimens.”[45] It isn’t possible to say in advance what kind of research may be performed at some point on stored biospecimens (such as blood samples and tissue biopsies), so this falls short of the usual standard for informed consent. Is such research nonetheless acceptable, given complete confidentiality and the fact that using the sample can’t physically harm the patient?

7.53 Randomization avoids bias. Suppose that you have 50 subjects available, 25 women and 25 men, to test reaction time to a visual cue in the presence or absence of a noise distraction. How does randomization distribute these subjects between the noise and no-noise groups? Start by labeling the women from 1 to 25 and the men from 26 to 50. Then use the *Simple Random Sample* applet to take 20 different random samples of size 25 from the 50 subjects. (Be sure to click “Reset” after each sample.) Record the counts of women (labels 1 to 25) in each of your 20 random samples. You should see that there is considerable chance variation but no systematic bias in favor of one or the

APPLET

other group in assigning the women. Larger samples from a larger population will, on average, do an even better job of creating two similar groups.

CHAPTER 8 Collecting and Exploring Data: Part I Review

Brigitte Baldi

IN THIS CHAPTER WE COVER...

- Part I Summary
- Comprehensive Review Exercises
- Large Data Set Exercises
- Online Data Sources
- EESEE Case Studies

Data analysis is the art of describing data using graphs and numerical summaries. The purpose of data analysis is to help us see and understand the most important features of a set of data. Chapter 1 described graphs to display distributions: pie charts and bar graphs for categorical variables, histograms and dotplots for quantitative variables, and time plots to show how a variable changes over time. Chapter 2 presented numerical tools for describing the center and spread of the distribution of one quantitative variable.

Figure 8.1 organizes the big ideas for exploring a quantitative variable. Plot your data, then describe their center and spread using either the mean and standard deviation or the five-number summary. The question marks in the figure's third box remind us that the usefulness of numerical summaries depends on what we find when we examine graphs of our data. No short summary does justice to irregular shapes or to data with several distinct clusters.

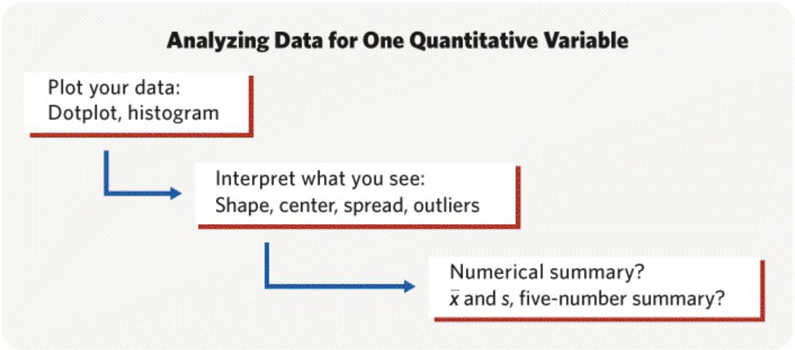

Analyzing Data for One Quantitative Variable
Plot your data:
Dotplot, histogram
Interpret what you see:
Shape, center, spread, outliers
Numerical summary?
$\bar{x}$ and s, five-number summary?

Figure 8.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 8.1

Chapters 3 and 4 applied the same ideas to relationships between two quantitative variables. Figure 8.2 retraces the big ideas of Figure 8.1, with details that fit the new setting. Always begin by making graphs of your data. In the case of a scatterplot, we have learned of a numerical summary relevant only for data that show a roughly linear pattern on the scatterplot. The summary for such data consists of the means and standard deviations of the two variables and their correlation. A regression line drawn on the plot gives a compact description of the overall pattern that we can use for prediction. The question marks in the figure remind us that correlation and regression describe only straight-line relationships.

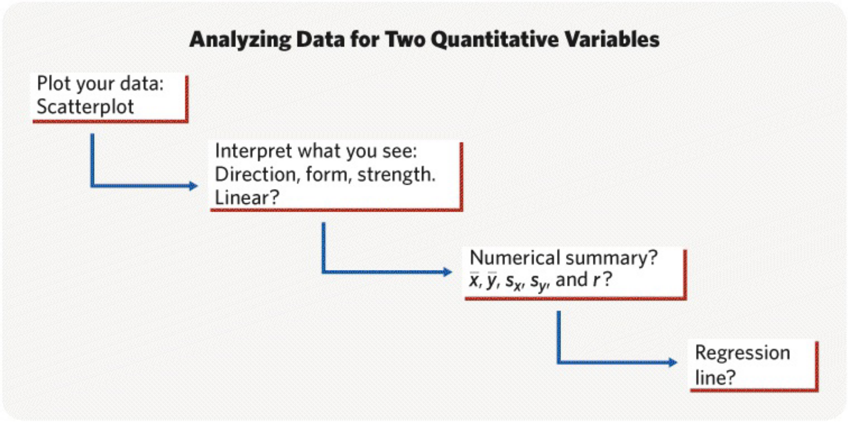

Analyzing Data for Two Quantitative Variables
Plot your data: Scatterplot
Interpret what you see: Direction, form, strength. Linear?
Numerical summary? $\bar{x}$, $\bar{y}$, s_x, s_y, and r?
Regression line?

Figure 8.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 8.2

Chapter 5 illustrated how to understand relationships between two categorical variables. Figure 8.3 displays the big ideas visually. First summarize the data as counts in a two-way table, then compare well-chosen percents. Marginal distributions describe only one categorical variable at a time. Conditional distributions, in contrast, describe the distribution of one variable for a specific outcome of the other variable. Therefore there are two different conditional distributions that can be used to examine the relationship between two categorical variables, and it is important not to confuse them when interpreting the data in context. When one variable is clearly the explanatory variable, the conditional distribution of the response variable given each level of the explanatory variable is typically more informative.

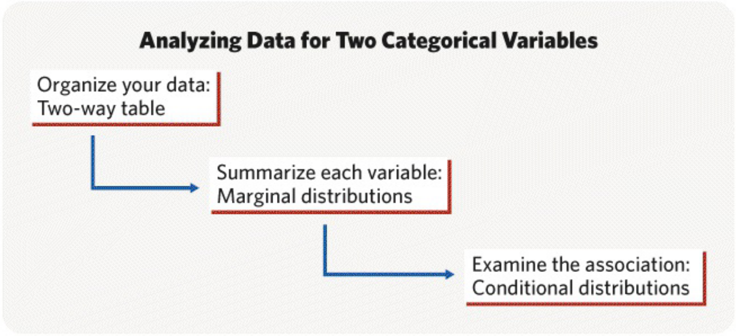

Analyzing Data for Two Categorical Variables
Organize your data: Two-way table
Summarize each variable: Marginal distributions
Examine the association: Conditional distributions

Figure 8.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 8.3

Chapters 6 and 7 showed us that knowing how the data were collected is essential to interpreting summary statistics in context. Figures 8.4 and

8.5 display the big ideas visually. Random sampling and randomized comparative experiments are perhaps the most important statistical inventions of the 20th century. The deliberate use of chance in producing data is a central idea in statistics. It not only reduces bias, but also allows use of the laws of probability to analyze data. You should now understand what are good techniques for producing data and why bad techniques often produce worthless data.

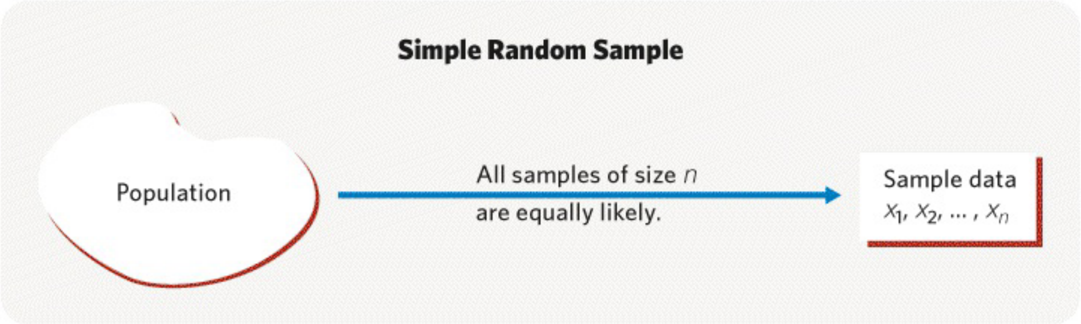

Simple Random Sample
Population
All samples of size n
are equally likely.
Sample data
$x_1, x_2, \ldots, x_n$

Figure 8.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 8.4

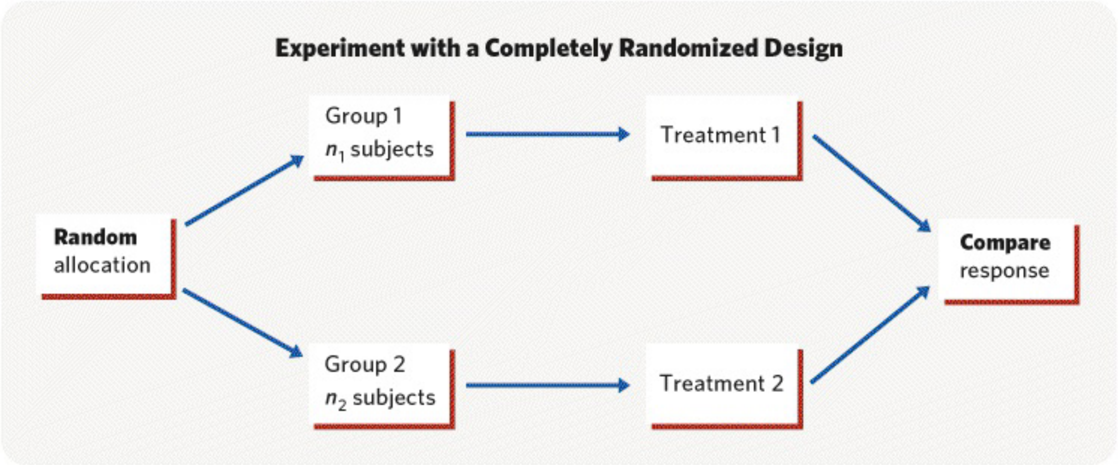

Experiment with a Completely Randomized Design
Group 1
n_1 subjects
Treatment 1
Random
allocation
Compare
response
Group 2
n_2 subjects
Treatment 2

Figure 8.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 8.5

You can organize your work in any open-ended data analysis setting by following the four-step "State, Plan, Solve, and Conclude" process first introduced in Chapter 2. After we have mastered the extra background needed for statistical inference, this process will also guide our practical work on inference later in the book.

PART I SUMMARY

Here are the most important skills you should have acquired from reading Chapters 1 to 7.

A. Data

1. Identify the individuals and variables in a set of data.
2. Identify each variable as categorical or quantitative. Identify the units in which each quantitative variable is measured.
3. Identify the explanatory (independent) and response (dependent) variables in situations where one variable explains or influences another.

B. Displaying Distributions

1. Recognize when a pie chart can and cannot be used.
2. Make a bar graph of the distribution of a categorical variable, or in general to compare related quantities.
3. Interpret pie charts and bar graphs.
4. Make a time plot of a quantitative variable recorded over time. Recognize patterns such as trends and cycles in time plots.
5. Make a histogram of the distribution of a quantitative variable.
6. Make a dotplot of the distribution of a small set of observations.
7. Describe the shape, center, and spread of a quantitative distribution from a histogram or a dotplot. Identify outliers of the distribution.

C. Describing Distributions (Quantitative Variable)

1. Look for the overall pattern and for major deviations from the pattern.
2. Assess from a histogram or dotplot whether the shape of a distribution is roughly symmetric, distinctly skewed, or neither. Assess whether the distribution has one or more major peaks.
3. Describe the overall pattern by giving numerical measures of center and spread in addition to a verbal description of shape.

4. Decide which measures of center and spread are more appropriate: the mean and standard deviation (especially for symmetric distributions) or the five-number summary (especially for skewed distributions).
5. Recognize outliers and give plausible explanations for them.

D. Numerical Summaries of Distributions (Quantitative Data)

1. Find the median MM and the quartiles Q1Q_1 and Q3Q_3 for a set of observations.
2. Find the five-number summary and draw a boxplot; assess center, spread, symmetry, and skewness from a boxplot.
3. (Optional) Use the 1.5×IQR$1.5 \times IQR$ rule to identify suspected outliers.
4. Find the mean x¯$\bar{x}$ and the standard deviation ss for a set of observations.
5. Understand that the median is more resistant than the mean. Recognize that skewness in a distribution moves the mean away from the median and toward the long tail.
6. Know the basic properties of the standard deviation: s≥0$s \geq 0$ always; s=0$s = 0$ only when all observations are identical and increases as the spread increases; ss has the same units as the original measurements; ss is increased by the presence of outliers or skewness.

E. Scatterplots and Correlation

1. Make a scatterplot to display the relationship between two quantitative variables measured on the same subjects. Place the explanatory variable (if any) on the horizontal scale of the plot.
2. Add a categorical variable (when available) to a scatterplot by using a different plotting symbol or color.
3. Describe the direction, form, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot and give plausible explanations for them.
4. Judge whether it is appropriate to use correlation to describe the

relationship between two quantitative variables. Find the correlation r.r.

5. **5.** Know the basic properties of correlation: rr measures the direction and strength of only straight-line relationships; rr is always a number between -1-1 and 1; r=±1$r = \pm 1$ only for perfect straight-line relationships; rr moves away from 0 and toward ±1± 1 as the straight-line relationship gets stronger.

F. Regression Lines

1. Understand that regression requires an explanatory variable and a response variable. Use a calculator or software to find the least-squares regression line of a response variable yy on an explanatory variable xx from data.
2. Explain what the slope bb and the intercept aa mean in the equation y^=a+bx$\hat{y} = a + bx$ of a regression line.
3. Draw a graph of a regression line when you are given its equation.
4. Use a regression line to predict yy for a given x.x. Recognize extrapolation and be aware of its dangers.
5. Find the slope and intercept of the least-squares regression line from the means and standard deviations of xx and yy and their correlation.
6. Use r2,r^2, the square of the correlation, to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.
7. Recognize outliers and potentially influential observations from a scatterplot with the regression line drawn on it.

G. Cautions about Correlation and Regression

1. Understand that both rr and the least-squares regression line can be strongly influenced by a few extreme observations.
2. Recognize possible lurking variables that may explain the observed association between two variables xx and y.y.
3. Understand that even a strong correlation does not mean that there is a cause-and-effect relationship between xx and y.y.

4. Give plausible explanations for an observed association between two variables: direct cause and effect, the influence of lurking variables, or both.

H. Relationships in Categorical Data

1. From a two-way table of counts, find the marginal distributions of both variables by obtaining the row sums and column sums.
2. Express any distribution in percents by dividing the category counts by their total.
3. Describe the relationship between two categorical variables by computing and comparing percents. Often this involves comparing the conditional distributions of one variable for the different categories of the other variable.
4. Recognize Simpson’s paradox and be able to explain it.

I. Sampling and Observational Studies

1. Identify the population in a sampling situation.
2. Identify potential confounding variables and understand the limitations of the conclusions that can be drawn from observational studies.
3. Recognize bias due to voluntary response samples and other inferior sampling methods.
4. Recognize the presence of undercoverage and nonresponse as sources of bias in a sample survey. Recognize the effect of the wording of questions on the responses.
5. Recognize designs commonly used in comparative observational studies, such as case-control and cohort designs.

J. Experiments

1. Recognize whether a study is an observational study or an experiment.
2. Recognize bias due to confounding of explanatory variables with lurking variables in either an observational study or an experiment.
3. Identify the factors (explanatory variables), treatments, response

variables, and individuals or subjects in an experiment.

4. Outline the design of a completely randomized experiment using a diagram like that in Figure 8.5. The diagram in a specific case should show the sizes of the groups, the treatments imposed, and the response variable.
5. Use software or a table of random digits to carry out the random assignment of subjects to groups in a completely randomized experiment.
6. Recognize the placebo effect. Recognize when the double-blind technique should be used.
7. Explain why randomized comparative experiments can give good evidence for cause-and-effect relationships.

COMPREHENSIVE REVIEW EXERCISES

These exercises are intended to help you integrate the concepts and approaches presented in Chapters 1 through 7.

8.1 Weight of newborns. Here is the distribution of the weight at birth for all babies born in the United States in 2010:[1]

Weight (grams)	Count	Weight (grams)	Count
Less than 500	5,980	3,000 to 3,499	1,566,755
500 to 999	22,015	3,500 to 3,999	1,055,004
1,000 to 1,499	29,846	4,000 to 4,499	262,997
1,500 to 1,999	63,427	4,500 to 4,999	36,706
2,000 to 2,499	204,295	5,000 to 5,499	4,216
2,500 to 2,999	744,181		

a. Display the data in a histogram, using the class *percents*, rather than the counts, on the vertical scale. Displaying relative frequencies makes it generally easier to compare data from different years or different countries. Describe the distribution of newborn weights.

b. In which weight class does the median birth weight fall?

c. Low-birth-weight babies, defined as babies born weighing at least 1500 grams (about 3 pounds 4 ounces) but less than 2500 grams (about 5 pounds 8 ounces), are at an increased risk of serious health problems. Very-low-birth-weight babies (weighing less than 1500 grams at birth) have the highest risk for health problems. Obtain the proportions of newborns who are normal-weight, low-birth-weight, and very-low-birthweight. Display these values in a pie chart or a bar graph.

d. Write a few sentences about birth weights in the United States based on your analysis.

8.2 Trends in manatee deaths. Examples 3.3 and 4.3 (pages 70 and 100) describe a data set from Florida relating manatee deaths from collisions with powerboats and the number of powerboats registered in a given year.

The data are available in Table 3.1 (page 70).

a. Complete the analysis of this data set by plotting and describing the trends over time for these two variables.
b. Are the data from an observational study or an experiment? What type of data collection design was used?
c. Write a short paragraph about manatee deaths from collisions with powerboats in Florida, combining the findings from your time-series analysis and the regression analysis from Example 4.3.

8.3 Paw preference in tree shrews. Tree shrews, *Tupaia belangeri*, are small omnivorous mammals phylogenetically related to primates. A research team examined paw preference among tree shrews born and raised in captivity. They computed a “pawedness index” (PI), a unitless continuous variable taking values between −1 and 1, depending on the relative use of the right and left paws during food-grabbing tasks. An animal always using its left paw would have a PI value of −1, but an animal using its right and left paws equally would have a PI value of 0. Here are the PI values for the 36 tree shrews, sorted for convenience:[2]

−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00	−1.00
−0.96	−0.95	−0.93	−0.89	−0.89	−0.25	−0.11	−0.08	−0.06	−0.03
0.13	0.24	0.41	0.50	0.56	0.67	0.86	0.87	0.90	0.94
1.00	1.00	1.00	1.00	1.00	1.00				

a. Display the PI values in a dotplot and use it to describe the distribution of PI values.
b. Explain why it would be misleading to report the mean and standard deviation for this data set. The five-number summary would provide more information, but it would still not be an appropriate summary description of this data set. You should always plot your data before computing numerical summaries.
c. The researchers labeled tree shrews with a pawdness index between −0.35 and 0.35 as ambidextrous. Shrews with values less than −0.35 were labeled left-biased and shrews with values greater than 0.35 were labeled right-biased. Obtain the proportions of right-biased, left-biased, and ambidextrous shrews in the study and display these values in a pie

chart or a bar graph.

8.4 Marijuana use among adolescents. The Substance Abuse and Mental Health Services Administration reports "past month marijuana use among adolescents in the United States" between 2002 and 2014 based on three different national surveys:[3] Monitoring the Future (MTF, 9th through 12th grades), the National Survey on Drug Use and Health (NSDUH, ages 12 through 17), and the Youth Risk Behavior Survey (YRBS, 8th, 10th, and 12th grades combined; conducted every other year). Here are the data taken from the published report.

Year	MTF	NSDUH	YRBS	Year	MTF	NSDUH	YRBS
2002	15.3	8.2		2009	13.8	7.4	20.8
2003	14.8	7.9	22.4	2010	14.8	7.4	
2004	13.6	7.6		2011	15.2	7.9	23.1
2005	13.4	6.8	20.2	2012	15.1	7.2	
2006	12.5	6.7		2013	15.6	7.1	23.4
2007	12.4	6.7	19.7	2014	14.4	7.4	
2008	12.5	6.7					

a. Create one time plot displaying all three data series. Describe the overall pattern and any notable similarities and differences between the three surveys.

b. In what ways does the data collection of these three surveys differ? Explain how these differences may contribute to differences in their findings.

c. Either one of these surveys may be used by researchers or news outlets as the sole source to generally describe trends in marijuana use among adolescents in the United States. Use your analysis to write a short paragraph explaining why details of a survey's methodology matter when reporting findings.

8.5 Falling through the ice. The Nenana Ice Classic is an annual contest to guess the exact time in the spring thaw when a tripod erected on the frozen Tanana River near Nenana, Alaska, will fall through the ice. The contest has been run since 1917. The 2017 jackpot was $267,444. Table

8.1 gives simplified data that record only the date on which the tripod fell each year. The earliest date so far is April 20. To make the data easier to use, the table gives the date each year in days starting with April 20. That is, April 20 is 1, April 21 is 2, and so on.[4]

TABLE 8.1 Days from April 20 for the Tanana River tripod to fall

Year	Day	Year	Day	Year	Day	Year	Day	Year	Day	Year	Day
1917	11	1934	11	1951	11	1968	19	1985	23	2002	18
1918	22	1935	26	1952	23	1969	9	1986	19	2003	10
1919	14	1936	11	1953	10	1970	15	1987	16	2004	5
1920	22	1937	23	1954	17	1971	19	1988	8	2005	9
1921	22	1938	17	1955	20	1972	21	1989	12	2006	13
1922	23	1939	10	1956	12	1973	15	1990	5	2007	8
1923	20	1940	1	1957	16	1974	17	1991	12	2008	17
1924	22	1941	14	1958	10	1975	21	1992	25	2009	12
1925	16	1942	11	1959	19	1976	13	1993	4	2010	10
1926	7	1943	9	1960	13	1977	17	1994	10	2011	15
1927	23	1944	15	1961	16	1978	11	1995	7	2012	4
1928	17	1945	27	1962	23	1979	11	1996	16	2013	31
1929	16	1946	16	1963	16	1980	10	1997	11	2014	6
1930	19	1947	14	1964	31	1981	11	1998	1	2015	5
1931	21	1948	24	1965	18	1982	21	1999	10	2016	4
1932	12	1949	25	1966	19	1983	10	2000	12	2017	12
1933	19	1950	17	1967	15	1984	20	2001	19		

a. Plot and describe the distribution of breakup dates, and provide appropriate numerical summaries.

b. If the date when the tripod falls has been getting earlier, that may be evidence for the effects of global warming. Create a time plot of the date the tripod falls against year. Describe the overall pattern. What does it suggest?

c. Further describe the pattern of fall dates as a function of time with a regression analysis. Obtain the equation of the least-squares regression line and indicate how much of the year-to-year variation in ice breakup time is explained by this trend over time.

8.6 Aging and the blood–brain barrier. The blood–brain barrier (BBB) is a structure found around blood vessels serving the brain that limits which substances may enter the brain, thereby providing protection from toxins and pathogens. Researchers recruited 24 healthy adults of varying ages with no cognitive disability and measured BBB permeability (in 10-3 $^{-3}$/min) using a high-resolution spatiotemporal imaging technique. Here are the findings for the hippocampus, a brain region involved in learning and memory:[5]

Age	Permeability	Age	Permeability	Age	Permeability
29	0.53	23	1.02	72	1.39
28	0.81	58	1.04	84	1.40
30	0.83	40	1.05	66	1.43
75	0.88	75	1.20	91	1.43
76	0.98	65	1.85	80	1.48
59	1.12	75	1.28	90	1.51
55	1.09	61	1.32	78	1.19
47	1.08	80	1.36	78	1.22

a. Identify the individuals, the response variable, and the type of study.

b. Plot and describe the distribution of BBB permeability, and provide appropriate numerical summaries.

c. Some cognitive functions, such as learning and memory, typically decline with age. Examine the distribution of BBB permeability as a function of age (shown here in years). Plot and describe their relationship, and provide appropriate numerical summaries. What percent of the variation in BBB permeability is explained by age?

d. Based on your analysis, explain why simply reporting summary statistics for BBB permeability in healthy adults without cognitive impairment would not provide an adequate benchmark to assess the neurovascular health of a patient with no cognitive impairment.

8.7 A big-toe problem. Hallux abducto valgus (call it HAV) is a deformation of the big toe that is not common in youth and often requires surgery. Metatarsus adductus (call it MA) is a turning in of the front part of the foot that is common in adolescents and usually corrects itself.

Doctors used X-rays to measure the HAV and MA angles of deformity (in degrees) in 38 consecutive patients younger than age 21 who came to a medical center for surgery to correct HAV.[6] More pronounced angles are indicative of a more serious deformity. The data appear in Table 8.2.

TABLE 8.2 **Angle of deformity (degrees) for two types of foot deformity**

HAV angle	MA angle	HAV angle	MA angle	HAV angle	MA angle
28	18	21	15	16	10
32	16	17	16	30	12
25	22	16	10	30	10
34	17	21	7	20	10
38	33	23	11	50	12
26	10	14	15	25	25
25	18	32	12	26	30
18	13	25	16	28	22
30	19	21	16	31	24
26	10	22	18	38	20
28	17	20	10	32	37
13	14	18	15	21	23
20	20	26	16		

a. Plot and describe the distribution of HAV angles among young patients needing surgery for this condition. Provide an appropriate numerical summary. If there are outliers, also obtain the summary statistics omitting them and discuss their influence.

b. Plot and describe the distribution of MA angles among young patients needing surgery for this condition. Provide an appropriate numerical summary. If there are outliers, also obtain the summary statistics omitting them and discuss their influence.

c. Doctors speculate that the severity of MA can help predict the severity of HAV. Make a scatterplot of the data and describe the relationship between MA angle and HAV angle, including any possible outliers. If appropriate, obtain the equation of the least-squares regression line for predicting HAV angle from MA angle.

d. Do the data appear to support the doctors' speculation? If so, how good a predictor would MA angle be? Provide a numerical measure to support your conclusion.

8.8 Myopia in children. Myopia, also called short-sightedness, is a common visual impairment that is reaching alarming proportions worldwide. The condition typically develops in school-age children and adolescents. It has long been thought that the condition was associated with too much reading, but recent evidence suggests instead that a lack of exposure to natural outdoor light may play a role. To investigate this new theory, researchers compared rates of myopia among randomly selected young children (6 and 7 years old) of Chinese ethnicity living in either Sydney (Australia) or Singapore. The researchers found that 4 of the 124 children in the Australian sample had myopia, compared with 183 among the 628 children from Singapore.[7]

a. Display the findings in a two-way table. What percent of the children in the study had myopia? What is that percent among the children living in Australia and among those living in Singapore?

b. The researchers obtained information for most of the children in the study about their weekly amount of time (in hours) spent reading or writing and the weekly amount of time (in hours) spent on outdoor activities. Here are the summary statistics:

Time reading/writing	**Number of children**	**Mean**	**Standard deviation**
Australia	109	20.8	13.9
Singapore	611	17.8	8.8

Time outdoors	**Number of children**	**Mean**	**Standard deviation**
Australia	102	13.75	1.02
Singapore	586	3.05	0.12

How do the children from both countries in this study differ in how they spend their time?

c. Does the study suggest that more time spent reading or writing is associated with a greater risk of myopia among the enrolled children? Does it suggest that more time spent outdoors is associated with a greater risk of myopia? Can we draw a conclusion of causality based on this study?

d. The children enrolled in this study were all of Chinese ethnicity. What is the advantage in this study of collecting data from children with the same ethnicity compared with selecting random samples of children from each location regardless of ethnicity?

8.9 Monkey calls. The usual way to study the brain's response to sounds is to have subjects listen to "pure tones." The response to recognizable sounds may differ. To compare responses, researchers anesthetized macaque monkeys. They fed pure tones and also monkey calls directly into their brains by inserting electrodes. Response to the stimulus was measured by the firing rate (electrical spikes per second) of neurons in various areas of the brain. Table 8.3 contains the responses for 37 neurons.[8]

TABLE 8.3 Neuron response (spikes per second) to tones and monkey calls

Neuron	Tone	Call	Neuron	Tone	Call	Neuron	Tone	Call
1	474	500	14	145	42	26	71	134
2	256	138	15	141	241	27	68	65
3	241	485	16	129	194	28	59	182
4	226	338	17	113	123	29	59	97
5	185	194	18	112	182	30	57	318
6	174	159	19	102	141	31	56	201
7	176	341	20	100	118	32	47	279
8	168	85	21	74	62	33	46	62
9	161	303	22	72	112	34	41	84
10	150	208	23	20	193	35	26	203
11	19	66	24	21	129	36	28	192
12	20	54	25	26	135	37	31	70
13	35	103						

a. Plot and describe the distribution of neuronal responses to a pure tone.

Provide an appropriate numerical summary. If there are outliers, also obtain the summary statistics omitting them and discuss their influence.

b. Plot and describe the distribution of neuronal responses to a monkey call. Provide an appropriate numerical summary. If there are outliers, also obtain the summary statistics omitting them and discuss their influence.
c. In what percent of the 37 neurons examined is the response to monkey calls stronger than the response to pure tones?
d. Is there a relationship between neuronal tone response and call response? Make a scatterplot using pure-tone response for the horizontal axis. Describe this relationship and, if appropriate, provide a measure of its strength.

8.10 Arthroscopic surgery. Arthroscopic surgery for the treatment of knee pain is one of the most common orthopedic procedures, despite a lack of relevant scientific studies of its efficacy. One recent, double-blind study randomly assigned patients with knee pain to arthroscopic surgery or sham surgery. One year later, 62 of the 70 patients in the treatment group reported improvement, compared with 63 of the 76 patients in the control group.[9]

a. Identify the individuals, the factors, the response variable, and the design used for this study.
b. Explain the terms “double-blind” and “sham surgery” in the context of this study.
c. Organize the study outcomes in a two-way table of counts. Obtain the conditional distribution of outcome, given treatment.
d. The difference between the proportions of patients who perceived improvement one year after undergoing either arthroscopic surgery or sham surgery is not statistically significant. Explain briefly what this means in the context of this problem.

8.11 The intriguing cerebellum. The cerebellum is a highly convoluted brain structure sitting underneath the cerebral hemispheres. This structure is thought to facilitate the acquisition and use of sensory data by the rest of the brain, particularly the motor areas. Studies suggest that the cerebellum may scale up with the size of animals' bodies and brains, whereas other parts of the brain are clearly represented differentially across species. Here are data on 15 mammal species showing the weights in grams for their bodies and cerebellums, along with the logarithm-transformed values:[10]

Species	**Body (g)**	**Cerebellum (g)**	**Log(body)**	**Log(cerebellum)**
Mouse	58	0.09	1.76	−1.05
Bat	30	0.09	1.48	−1.05
Flying fox	130	0.30	2.11	−0.52
Pigeon	500	0.40	2.70	−0.40
Guinea pig	485	0.90	2.69	−0.05
Squirrel	350	1.50	2.54	0.18
Chinchilla	500	1.70	2.70	0.23
Rabbit	1,800	1.90	3.26	0.28
Hare	3,000	2.30	3.48	0.36
Cat	3,500	5.30	3.54	0.72
Dog	3,500	6.00	3.54	0.78
Macaque	6,000	7.80	3.78	0.89
Sheep	25,000	21.50	4.40	1.33
Bovine	300,000	35.70	5.48	1.55
Human	60,000	142.00	4.78	2.15

a. Display the body weights and cerebellum weights in separate dotplots and describe their distributions.

b. Display the logarithm transformed body weights and cerebellum weights in separate dotplots and describe their distributions. How did the logarithm transformation affect the shape of the distributions of these two variables? This effect is the reason why strongly right-skewed variables are sometimes expressed in a logarithm scale.

c. Make a scatterplot of cerebellum weight (yy) against body weight (xx) and another scatterplot using the transformed values. Describe and compare both scatterplots. What did you learn about the relationship between cerebellum weight and body weight from these graphs? Do these data appear to support the idea that the cerebellum scales up with the size of animals' bodies and brains?

d. Add the least-squares regression line to the log-transformed scatterplot. In the logarithm scales, what proportion of the variation in cerebellum weight can be explained by the variation in body weight?

8.12 A vaccine–autism link? One particular research study published in 1998 described a link between vaccination (specifically, the measles, mumps, rubella [MMR] vaccine) and autism. The publication received a lot of attention from the news media, in part instigated by its principal author, Dr. Wakefield. The news led to worldwide concerns over the safety of vaccines and drastic decreases in vaccination rates.

a. Go to http://dx.doi.org/10.1016/S0140-6736(97)11096-0 to access the publication. Reading the abstract is sufficient for this exercise, though the entire article is also accessible. Describe the study. Was its design observational or experimental? Who were the individuals studied? How many were there and how were they selected? What is the strength of the evidence presented here for a causal link between MMR vaccination and autism in children?

b. This publication (and following press interviews) did not disclose that the patients had been recruited through anti-MMR campaigners, that the study had been commissioned and funded for the purpose of a lawsuit, and that the principal author was a paid consultant for this litigation. The publication did not disclose either that Dr. Wakefield had filed patents in direct conflict of interest with the existing MMR vaccine. Explain why it is important that such information be disclosed when reporting a study's findings. Which part of the study is especially affected by this undisclosed information?

c. An investigative reporter later uncovered systematic data falsification in the 1998 publication.[11] Most notably, the timing of first autistic symptoms was shortened from months to days for some children and it was even reversed for some of the other children (these children actually had diagnosed symptoms before being vaccinated). Also, claims that the lab tests showed inflammation of the colon were actually made up. Since 1998, numerous studies have been conducted worldwide and none has found a link between autism and the MMR vaccine or any of its components.[12] Describe the strength of the evidence for a causal link between MMR vaccination and autism in children in light of this additional information.

LARGE DATA SET EXERCISES

These exercises guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. They cover concepts and approaches presented in Chapters 1 through 7. Note that short answers are not available for these exercises to enable their use for class work or assignments.

8.13 Survey data. The data file *Large.Survey* contains the anonymous answers of students in a large statistics class to 5 simple questions. The point here is to illustrate some of the challenges one can encounter with data collection.

LARGE DATA SET

a. The first step is getting the data ready for analysis. The survey was administered online and the program used generated column headers using the questions asked. Start by giving each column a simple variable name: blood, height, sex, weight, eye. Then examine the data for potential issues. Students were instructed to enter numerical values in a specific format, which they did not all follow to the letter. Because the data were collected anonymously, there is no way to investigate these problematic values (although some errors are easy enough to correct).There are several approaches to finding problematic data. Some statistical software will let you use conditional formatting to highlight values you deem too large or too small to be realistic. Plotting the data also helps find outliers that are obvious errors. Write a short summary of the problems you noticed and how you addressed them.

b. Two of the variables are categorical. The survey offered a list of choices for students to select. There are therefore no possible typos in these answers and no problems with similar answers recorded differently (for example, “brown” versus “Brown”). Plot the distribution of blood types using a pie chart or a bar graph. Apparently, a lot of students either did not know their blood type or chose not to provide that information. Plot the distribution of blood types again, this time ignoring the “I don’t remember” answers. Describe your findings. Plot and describe the distribution of eye color.

c. There are two quantitative variables. These are morphological variables known to differ noticeably between males and females. Plot the distribution of heights for males and females separately, using a histogram or a dotplot. Do the same for the distribution of weights. Obtain appropriate numerical summaries, and then write a few sentences describing your findings.

d. Examine the relationship between height (explanatory) and weight, using different symbols for males and for females. Describe these relationships and give a numerical summary of their strength.

8.14 Essential tremor. Essential tremor is neurological movement disorder characterized by involuntary rhythmic movement, which typically interferes with the full use of the arms and hands. The most common movement disorder, it differs from Parkinson’s disease in that it is typically action related. Researchers recruited a sample of 15 subjects diagnosed with essential tremor to test the efficacy of a tremor-canceling device. Before proceeding with the testing, the researchers obtained the baseline characteristics of the participating subjects.[13] Some of the findings are available in the *Large.Tremor* data file. “TRS” stands for Tremor Rating Scale, a measure of tremor severity.

LARGE DATA SET

a. Plot and describe the distribution of TRS total scores, making sure to provide an appropriate numerical summary.

b. Essential tremor can occur at any age (though it is most common in people older than age 40) and the condition often worsens over time. Investigate the relationship between TRS total score and disease duration in this sample of essential tremor patients.

c. Investigate whether your previous conclusions are affected by the patient’s sex.

8.15 Butterfly wings. Monarch butterflies (*Danaus plexippus*) undertake large seasonal migrations over North America. Studies indicate that the females of the species may have a higher migratory success and flight ability. Researchers studied the sex-based morphological differences of monarch butterflies related to flight. They measured the weight (in mg) and surface area (in mm^2) of a sample of 92 monarch butterflies, all of which were reared in captivity in identical conditions.[14] The data are available in the *Large.Butterfly* data file.

a. Plot and describe the distribution of forewing weights separately for the male (*Female* = 0) and female (*Female* = 1) butterflies. Look for potential outliers and provide appropriate numerical summaries.

b. Plot and describe the distribution of forewing surface areas separately for the male and female butterflies. Look for potential outliers and provide appropriate numerical summaries.

c. Examine the relationship between forewing surface area (explanatory) and weight separately for the male and female butterflies. Describe the relationships and compute appropriate numerical summaries for the relationships.

d. Write a short paragraph describing the morphological differences between male and female monarch butterflies, referencing your work.

e. The researchers concluded that “female wings are generally sturdier,” a finding “consistent with field observations made during the monarch migration, where wing condition (i.e., degree of damage and wear) of females has been found to be significantly better than those of male.” Which aspect of your own analysis supports this interpretation? Explain your answer.

8.16 Estimating body fat in men. The data file *Large.Bodyfat* contains data on the percent body fat, age, weight, height, and 10 body circumference measurements for 252 adult men.[15]

a. Make a histogram of the men’s body weights (in pounds). Describe the distribution and any deviations from the main pattern. Obtain the five-number summary and look for suspected outliers using the 1.5×IQR

$1.5 \times IQR$ rule. Which are actual outliers? There is one extreme outlier. Is this nonetheless a plausible value (that is, do individuals like that exist)? Check all the values given for that man before you answer.

b. Make a histogram of the men's heights (in inches). Describe the distribution and any deviations from the main pattern. Obtain the five-number summary and look for suspected outliers using the 1.5×IQR $1.5 \times IQR$ rule. Which are actual outliers? There is one extreme outlier. Is this nonetheless a plausible value (that is, do individuals like that exist)? This was actually a typo made in recording the man's height and this data point should be excluded when performing further analysis involving height.

8.17 More on estimating body fat in men. Continue your work from Exercise 8.16 with the data file *Large.Bodyfat.*

a. A man's weight depends on a number of factors, including his height. Plot and describe the relationship between height (explanatory) and weight (response). You previously identified two extreme outliers. Point them out on your scatterplot.

b. The study made two accurate measurements of the percent body fat (percent fat1 and percent fat2) based on two computation methods using underwater weighing. Make a scatterplot of the two variables (neither is explanatory here) and describe the relationship. Find the correlation coefficient r.r. Do the two methods give similar results?

c. Accurate measurements of the percent body fat require underwater weighing—a time-consuming and expensive process. We'd like to know if simpler body measurements can provide a good approximation of a person's percent body fat. Make a series of scatterplots with percent fat 1 on the yy axis(response)and each of the 10 body circumference variables, in turn, on the xx axis. If appropriate, calculate rr^2 in each case. Which of these 10 body circumference variables seems like the best candidate to use as a predictor for percent body fat?

d. Find the equation for the least-squares regression predicting percent fat1 from abdomen circumference. What percent of the variation in percent fat1 is explained by this model? What percent body fat would you predict for an adult man with an abdomen circumference of 100 cm?

8.18 Ant morphology. Researchers were interested in the epigenetic factors that may influence quantitative traits such as size in the carpenter ant (*Camponotus floridanus*).[16] The file *Large.Ant* contains data on the 179 adult carpenter ants (workers) they sampled from a mature ant colony: their scape length (in microns; a simple and common proxy for body size), and their head width (also in microns).

a. Plot and describe the distribution of scape lengths. Look for potential outliers and unusual patterns.
b. Plot and describe the distribution of head widths. Look for potential outliers and unusual patterns.
c. Morphology is defined by body proportions. Examine graphically the relationship between scape length (explanatory) and head width in adult carpenter ants. What does your analysis reveal?
d. The researchers determined that the observed variations in body proportions indicate that there are two types of workers among adult carpenter ants, which they labeled major workers and minor workers. Obtain appropriate numerical summaries of scape length and head width for each group. Create corresponding boxplots and write a short paragraph describing the morphological differences between major and minor carpenter ant workers, referencing your work.

8.19 Lung capacity of children. The forced expiratory volume (FEV, measured in liters) is a primary indicator of lung function and corresponds to the volume of air that can forcibly be blown out in the first second after full inspiration. The *Large.FEV* data file contains the FEV values of a large sample of children, along with some categorical descriptors of each individual.[17]

a. Make a histogram of FEV and describe the distribution.
b. Make separate histograms of FEV for girls (*sex* = 0) and for boys (*sex* = 1), using the same horizontal scale for both. Describe the differences between the two groups.
c. Create a new variable based on age: *preschool* for ages 3 to 5,

elementary for ages 6 to 10, *middle* for ages 11 to 13, and *highschool* for ages 14 and above. Then make separate histograms using the same horizontal scale for each of the four age groups. What do you notice?

d. Obtain summary statistics for each of the eight combinations of sex and age group. Make aside-by-side boxplot comparing these eight combinations. Does this last analysis change your previous conclusions? If so, explain how.

e. Write a short paragraph describing the differences between girls and boys by age group, referencing your work.

8.20 More on the lung capacity of children. Continue your work from Exercise 8.19 with the data file *Large.FEV.*

LARGE DATA SET

a. How is lung capacity affected by children’s growth? Plot the relationship between FEV and age for all the individuals in the data set, using FEV as the response variable. Describe the relationship. Do the same for the relationship between FEV and height.

b. Display the relationships in part a separately for boys and for girls (on separate graphs or on the same graph using different symbols). How do boys and girls differ in lung capacity as they age and grow?

c. Now examine the impact of smoking on lung capacity for children in middle and high school (ages 11 and up). Display the relationships in part a separately for smokers and nonsmokers for children age 11 and older. How does smoking affect the change of lung capacity as children grow and age?

d. Among the younger children (preschool and elementary school: ages 10 and younger), boys and girls are similar enough to be studied together. Plot the relationship between FEV and height for these younger children and obtain the least-squares regression equation. What percent of the variation in FEV is explained by this model? Predict the FEV of a child who is 5 feet tall (60 inches).

e. Our analysis so far suggests that, among adolescents, boys and girls should be studied separately. Plot the relationship between FEV and height for adolescent girls and obtain the least-squares regression equation. What percent of the variation in FEV is explained by this model? Predict the FEV of an adolescent girl who is 68 inches tall.

f. Plot the relationship between FEV and height for adolescent boys and obtain the least-squares regression equation. What percent of the variation in FEV is explained by this model? Predict the FEV of an adolescent boy who is 68 inches tall.

8.21 Diabetes in the United States. The Centers for Disease Control and Prevention (CDC) offers an interactive data visualization tool (www.cdc.gov/diabetes/atlas/obesityrisk/atlas.html) to examine diabetes trends in the United States. State age-adjusted rates of diabetes among adults (in percents) are shown on the left as a bar graph (bottom) and a map (top). The association between diabetes and obesity among adults is displayed in a scatterplot on the top right (the size of the circles is scaled to reflect the number of affected individuals in each state). You can view these data over time by selecting the timeline above the set of graphs. A time plot of the median diabetes rate among all states is displayed at the bottom right. Hovering over the graphs provides the exact values for each point and location and highlights the corresponding data in the other graphs. (Note that the data can also be downloaded for analysis with your own statistical software package.)

LARGE DATA SET

a. Examine the latest year. Describe the range of diabetes rates in the U.S. states and territories and report the median rate. Describe the relationship between obesity rates and diabetes rates. Can we conclude that obesity causes diabetes? Explain your reasoning.

b. Use the timeline to explore the change in diabetes rates in the United States over the past few decades. Describe the spatial and temporal trends you notice.

c. Use the timeline to explore the change in the association between obesity rates and diabetes rates in the United States over the past few decades. Describe both the differences and similarities in this association over time.

d. Pick one state or territory (if you live in the United States, select the state where you live). Write a short paragraph summarizing your findings for that state over time and describing how it contrasts with the rest of the states overall.

ONLINE DATA SOURCES

Here are a few regularly updated, online data sources relevant to the life sciences that can help further illustrate and practice the concepts introduced in Chapters 1 through 7. Those marked with an asterisk provide embedded data analysis and visualization tools.

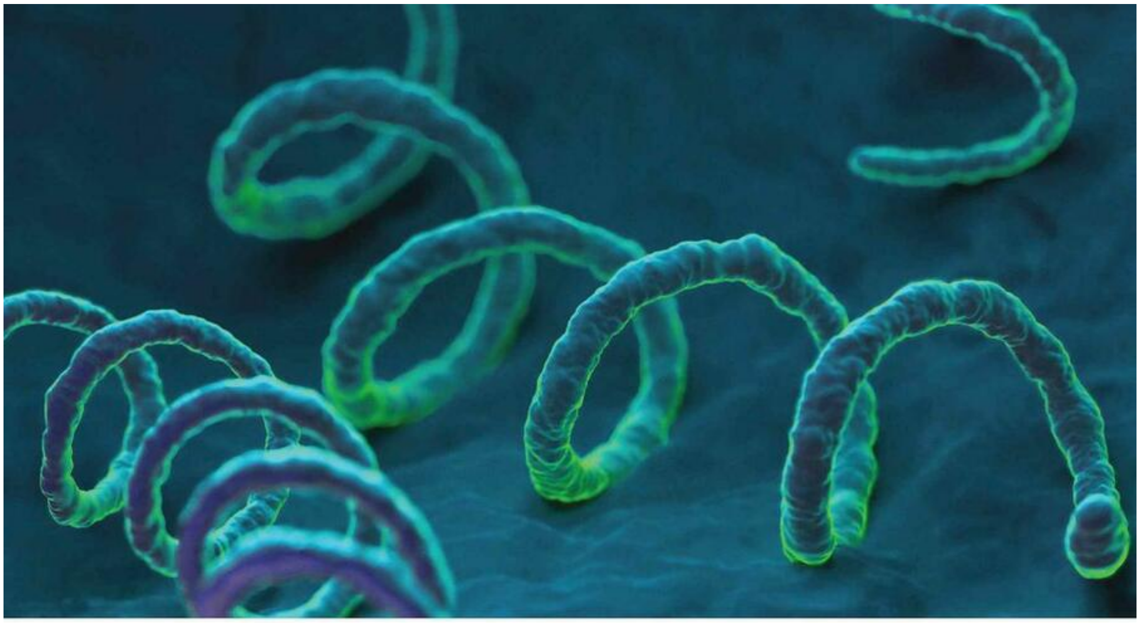
- blogs.cdc.gov/nchs-data-visualization * CDC, National Center for Health Statistics: Data Visualization Gallery
- www.cdc.gov/diabetes/atlas * CDC, Diabetes and Data Statistics: geographical and temporal trends in diabetes and associated risk factors
- chhs.data.ca.gov * California Health and Human Services: Open Data Portal
- nccd.cdc.gov/NPAO_DTM * CDC, Division of Nutrition, Physical Activity and Obesity: Data, Trends and Maps
- stateofobesity.org/rates * Trust for America's Health and Robert Wood Johnson Foundation: interactive data and reports
- www.cdc.gov/diabetes/statistics/complications_national.htm CDC: diabetes data and graphs
- www.cdc.gov/obesity/data/prevalence-maps.html CDC: obesity data and maps
- data.giss.nasa.gov/gistemp NASA, Goddard Institute for Space Studies: global data and graphs
- data.giss.nasa.gov/gistemp/station_data NASA, Goddard Institute for Space Studies: station data and graphs
- www.sciencedaily.com ScienceDaily: repository of newly published science articles provided with a summary in layman's terms and a link to the original source material

EESEE CASE STUDIES

The Electronic Encyclopedia of Statistical Examples and Exercises (EESEE) is available on the website. These more elaborate stories, with data, provide settings for longer case studies. Here are some suggestions for EESEE stories that apply the ideas you have learned in Chapters 1 through 7.

- **Acorn Size and Oak Tree Range.** Questions 1 and 2.
- **Anecdotes of Bias.** All questions.
- **Blinded Knee Doctors.** Questions 1, 2, and 3.
- **Brain Size and Intelligence.** Question 3.
- **Checkmating and Reading Skills.** Question 1.
- **Counting Calories.** Questions 1, 4, and 6.
- **Habitat of the Spotted Owl.** Questions 1 and 3.
- **Is Caffeine Dependence Real?** Questions 1, 2, and 3.
- **Keeping Balance as You Age.** Questions 1 and 4.
- **Mercury in Florida's Bass.** Question 5. (Note: "Homoscedastic" means that the scatter of points about the overall pattern is roughly the same from one side of the scatterplot to the other.)
- **Surgery in a Blanket.** Questions 1 and 2.
- **Visibility of Highway Signs.** Questions 1, 2, 3(a), and 5.

PART II



Science Picture Co./Getty Images

The purpose of statistics is to learn from data. In Part I we studied one approach to data, called *exploratory data analysis*. Data can also be described with mathematical models that follow a set of logic-based rules. This is the domain of *probability*.

We first examined the idea of chance in the context of designing studies. In Part II, Chapter 9 and the optional Chapter 10 lay out the formal language of probability and the logic rules it obeys. Chapter 11 and the optional Chapter 12 describe some important probability distributions and their main applications. Chapter 13 brings us back to data collection by using probability to describe how random samples vary.

Having a probability distribution to model the expected behavior of random samples allows us to draw conclusions about populations on the basis of random samples and to use probability to indicate how reliable the conclusions are. That's *statistical inference*:

- The purpose is to answer specific questions posed before the data were produced.

- Conclusions are formal, backed by a statement of our confidence in them, and they apply to a larger group of individuals or a broader class of circumstances.

Chapters 14 and 15 present the core concepts of *inference*. Chapter 16 reviews the material of Chapters 9 to 15 and provides more comprehensive exercises.

From Chance to Inference

PROBABILITY RULES

Chapter 9 Essential Probability Rules

Chapter 10 Independence and Conditional Probabilities*

PROBABILITY DISTRIBUTIONS

Chapter 11 The Normal Distributions

Chapter 12 Discrete Probability Distributions*

Chapter 13 Sampling Distributions

THE IDEA OF INFERENCE

Chapter 14 Introduction to Inference

Chapter 15 Inference in Practice

Chapter 16 From Chance to Inference: Part II Review

CHAPTER 9 Essential Probability Rules

Agency Animal Picture/Getty Images

IN THIS CHAPTER WE COVER...

- The idea of probability
- Probability models
- Probability rules
- Discrete versus continuous probability models
- Random variables
- Risk and odds

In Chapters 6 and 7, we saw that we can rely on chance—that is, random selection—to select samples without built-in biases so they can be used to infer the properties of a target population. In this and the following chapters of Part II, we explore the mathematics of chance behavior that are at the foundation of statistical inference. We also examine how probability *models* can be used to make predictions about events with uncertain outcomes.

THE IDEA OF PROBABILITY

Somewhere in the United States, the next baby is about to be born. Will this baby be a boy or a girl? We cannot know for sure because the outcome is not predetermined. But there is still a regular pattern that emerges clearly when many births are examined. This remarkable fact is the basis for the idea of probability. **Chance behavior is unpredictable in the short run but has a regular and predictable pattern in the long run.**

RANDOMNESS AND PROBABILITY

We call a phenomenon **random** if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.

EXAMPLE 9.1 Sex of newborns: theory and practice

A baby's sex depends on a pair of chromosomes, XX for girls and XY for boys, with each chromosome being inherited from one parent. Mendel's theory of genetic inheritance predicts that the XX and XY pairs should be roughly equally represented in the entire population of newborns (ova carry an X chromosome but spermatozoids carry either an X or a Y chromosome). Based on this scientific theory, we assign an equal chance that the next newborn will be a girl (probability 0.5) or a boy (probability 0.5).

In practice, other factors may affect the success rate of gametes and the relative survival of male and female embryos until birth. Here are some birth data for the United States in the past decades:[1]

	1990	1996	2002	2008	2014
Number of births	4,158,212	3,891,494	4,021,726	4,247,694	3,988,076

Proportion of males	0.5121	0.5115	0.5117	0.5117	0.5117

katrinaelena/Depositphotos

These values are reported by the U.S. National Center for Health Statistics “based on 100% of the birth certificates registered in all states and DC.” From these reports, we can assign a slightly different chance that the next newborn will be a girl (probability 0.488) or a boy (probability 0.512).

Whether we have a scientifically founded theory or empirical evidence for the entire population, we can use the long-term pattern to create a *probability model* to describe and quantify the uncertainty in random phenomena such as a newborn’s sex. Some models may be more precise than others, but they are still models: that is, they are mathematical representations of underlying patterns.

We may not always have a scientific theory or a census of an entire population to generate a sound probability model. However, in Chapters 6 and 7, we discussed how random samples and randomized experiments can be used to infer the properties of the populations from which they were drawn. Let’s look at a typical sample survey.

EXAMPLE 9.2 How common is the common cold?

The Gallup-Healthways sample survey collects health information about the American population. In February 2016, Gallup took a national random sample of 14,169 adults and found that 1587 (11.2%) of the people in the sample said they were sick with a cold the previous day.

Because all adults had the same chance to be among the chosen 14,169, it seems reasonable to use this 11.2% as an estimate of the unknown proportion in the population. It’s a *fact* that 11.2% of the sample were sick with a cold—we know because Gallup asked them. We don’t know what percent of all adults in the United States were sick with a cold in February 2016, but we *estimate* that approximately 11.2% were.

If we use the sample proportion of 11.2% as an approximate probability that a randomly selected American adult had a cold in February 2016, how do we know that it is a reasonable probability model?

What if Gallup had taken a second random sample of 14,169 adults and asked them the same question? The new sample would include different people. It is almost certain that there would not be exactly 1587 “sick with a cold” responses in this sample. That is, Gallup’s estimate of the proportion of adults who were sick with a cold in February 2016 will vary from sample to sample. Could it happen that one random sample finds that 11.2% of adults were sick with a cold and a second random sample finds that 15.3% were? *Random samples eliminate bias from the act of choosing a sample, but they can still be off because of the variability that results when we choose at random.* If the variation when we take repeat samples from the same population is too great, we can’t trust the results of any one sample. We will examine how much results from random samples can vary in Chapter 13.

CAUTION

For now, let’s approach this issue by considering what happens for a similar but easier (and cheaper) situation: coin tossing.

EXAMPLE 9.3 Coin tossing

When you toss a coin, there are only two possible outcomes, heads or tails. Figure 9.1 shows the results of simulating a coin toss 5000 times in a row on two separate occasions. After each toss from 1 to 5000, the proportion of all tosses so far that gave a head is recomputed and plotted on the vertical axis. (Notice that the horizontal axis is shown on a logarithmic scale to help us better see the initial tosses.) Trial A (solid line) begins tail, head, tail, tail. You can see that the proportion of heads for Trial A starts at 0 on the first toss, rises to 0.5 when the second toss gives a head, then falls to 0.33 and 0.25 as we get two more tails. Trial B (dotted line) starts with five straight heads, so the proportion of heads is 1 until the sixth toss.

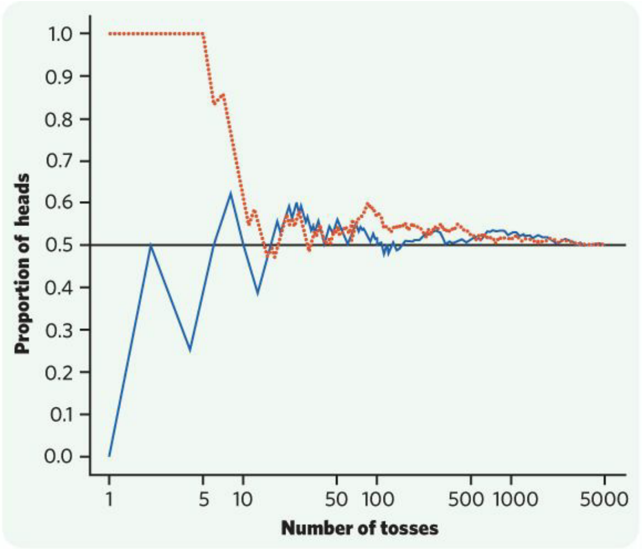


Figure 9.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 9.1 The proportion of tosses of a coin that give a head changes as we make more tosses. Eventually, however, the proportion approaches 0.5, the probability of a head. This figure shows the results of two trials of 5000 consecutive tosses each.

The proportion of tosses that produce heads is quite variable for a relatively small number of tosses. Trial A starts low and Trial B starts high. As we make more tosses, however, the proportion of heads for both trials gets close to 0.5 and stays there. If we conducted a third trial in which we tossed the coin a great many times, the proportion of heads would again settle down to 0.5 in the long run. This is the intuitive idea of probability. Probability 0.5 means “occurs half the time in a *very large* number of trials.” The probability 0.5 appears as a horizontal line on the graph.

In the Gallup survey described in Example 9.2, every adult interviewed answered either “sick with a cold” or “not sick with a cold.” This is similar to the “head” and “tail” outcomes of tossing a coin (except that the two outcomes are not equally likely). You can simulate random behavior as Figure 9.1 does by using the *Probability* applet. It allows you to choose the probability of a head and simulate any number of coin tosses with that probability. Experience shows that the proportion of heads gradually settles down close to the chosen probability as the number of tosses increases. Equally important, it also shows that *the proportion in a small or moderate number of tosses can be far from the actual probability*.

APPLET

CAUTION

Chapter 13 on sampling distributions will provide a more in-depth study of how random samples vary. And we will see that the expected deviation between the sample proportion and the true proportion in the population is smaller when we rely on larger samples. For now, we simply focus on the fact that the outcome of a *very large* random sample or randomized experiment can be used to compute *approximate* probabilities.

Computing probabilities based on what happens in the long run is called a **frequentist** approach to defining probabilities, because we rely on the relative frequency (proportion) of one particular outcome among very many observations of the random phenomenon. This idea ties probability to actual outcomes. Yet we often encounter another, quite different, idea of probability.

EXAMPLE 9.4 Intelligent life and the universe

Joe reads an article discussing the Search for Extraterrestrial Intelligence (SETI) project. We ask Joe, “What’s the chance that we will find evidence of extraterrestrial intelligence in this century?” Joe responds, “Oh, about 1%.”

Does Joe assign probability 0.01 to humans finding extraterrestrial intelligence this century? The outcome of the search is certainly unpredictable, but we can’t reasonably ask what would happen in many repetitions. This century will happen only once and will differ from all

> other centuries in many ways, especially in terms of technology. If probability measures “what would happen if we did this many times,” Joe’s 0.01 is not a probability. The frequentist definition of probability is based on data from many repetitions of the same random phenomenon. Joe is giving us something else—his personal judgment.

Although Joe’s 0.01 isn’t a probability in the frequentist sense, it gives useful information about Joe’s opinion that can be expressed in the language of probability. Such values are called *personal probabilities*.

PERSONAL PROBABILITY

A **personal probability** of an outcome is a number between 0 and 1 that expresses an individual’s judgment of how likely the outcome is.

STATS IN YOUR WORLD

WHAT LOOKS RANDOM?

Toss a coin six times and record heads (H) or tails (T) on each toss. Which of these outcomes is more probable: HTHTTH or TTTHHH? Almost everyone says that HTHTTH is more probable, because TTTHHH does not “look random.” In fact, both are equally probable. That heads has probability 0.5 says that approximately half of a very long sequence of tosses will be heads. It doesn’t say that heads and tails must come close to alternating in the short run. The coin doesn’t know what past outcomes were, and it can’t try to create a balanced sequence.

Personal probabilities are indeed personal: They vary from person to person. However, they are not necessarily arbitrary and flimsy. Joe may know nothing about the search for extraterrestrial life beyond what he learned during a quick perusal of the article, or he might be an expert whose opinion is based on conclusions from numerous rigorous scientific analyses. When scientists write, “It is *very likely* that the 21st-century mean rate of global mean sea level rise (...) will exceed that of 1971–2010,”[2] their stated opinion is grounded on extensive data collection and analysis, comprehensive computer simulations, and the consideration of various future global economic scenarios.

Regardless of their frequentist or personal interpretation, and regardless of the method used to derive them, all probabilities follow the probability rules we will describe in this and the following chapters.

APPLY YOUR KNOWLEDGE

9.1 Hemophilia. Hemophilia refers to a group of rare hereditary disorders of blood coagulation. Because the disorder is caused by defective genes on the X chromosome, hemophilia affects primarily men. According to the Centers for Disease Control and Prevention (CDC), the prevalence of hemophilia (the number in the population with hemophilia at any given time) among American males is 13 in 100,000. Explain carefully what this means. In particular, explain why it does *not* mean that if you obtain the medical records of 100,000 males, exactly 13 will be diagnosed with hemophilia.

9.2 Random digits. The table of random digits (Table A) was produced by a random mechanism that gives each value in the table probability 0.1 of being a zero.

a. What proportion of the first 50 digits in the table are 0s? This proportion is an estimate, based on 50 repetitions, of the true probability, which in this case is known to be 0.1.

b. The *Probability* applet can imitate random digits. Set the probability of heads in the applet to 0.1. Check “Show true probability” to show this value on the graph. A head stands for a 0 in the random digits table, and a tail stands for any other digit. Simulate 200 digits. If you kept going forever, presumably you would get 10% heads. What was the result of your 200 tosses?

9.3 Probability says … Probability is a measure of how likely an event is to occur. Match the probabilities that follow with each statement of likelihood given. (The probability value is usually a more exact measure of likelihood than is the verbal statement.)

0   0.01   0.6   0.99   1

a. This event is impossible. It can never occur.
b. This event is certain. It will occur on every trial.
c. This event is very unlikely, but it will occur once in a while in a long sequence of trials.
d. This event will occur more often than not.

9.4 Will you have an accident? The probability that a randomly chosen adult will be involved in a car accident in the next year is approximately 0.2. This probability is based on the proportion of millions of drivers who have accidents. “Accident” includes events such as crumpling a fender in your own driveway, not just highway accidents.

a. What do you think is your own probability of being in an accident in the next year? This is a personal probability.
b. Give some reasons why your personal probability might be a more accurate prediction of your “true chance” of having an accident than the probability for a random driver.
c. Almost everyone says that their personal probability is lower than the random driver probability. Why do you think this is true?

PROBABILITY MODELS

We’ve discussed the idea of probability in the context of a few examples. We now formalize the expression of probability models. In Example 9.1, we considered the sex of the next baby about to be born. We do not know in advance what the baby’s sex will be, but we can say that the outcome will be either male or female. Based on Mendelian genetic theory, we assign probability 1/2 to each of these outcomes. This description of a newborn’s sex has two parts:

- A list of possible outcomes
- A probability for each outcome

Such a description is the basis for all probability models. Here is the basic vocabulary we use.

PROBABILITY MODELS

The **sample space** SS of a random phenomenon is the set of all possible outcomes.

An **event** is an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.

A **probability model** is a mathematical description of a random phenomenon consisting of two parts: a sample space S S and a way of assigning probabilities to events.

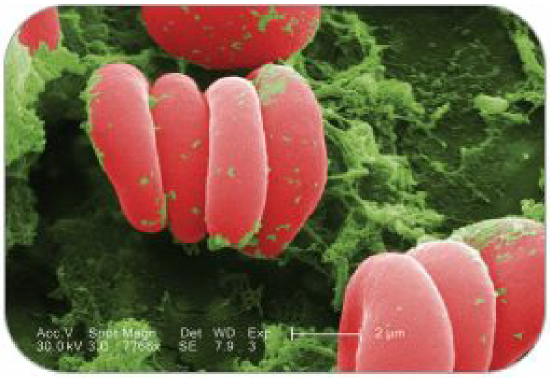
A sample space SS can be a simple list or a more comprehensive description. When one baby is born, there are only two outcomes, male and female. The sample space is S={M,F}.$S = \{M, F\}$. When the National Health Survey records the body weights in pounds of a random sample of adults, the sample space contains all possible adult weights over a realistic interval.

EXAMPLE 9.5 Blood types

Your blood type greatly impacts the kind of blood transfusion or organ transplant you can safely get. Eight different blood types occur as a result of the presence or absence of certain molecules on the surface of red blood cells. A person’s blood type is given as a combination of a group (O, A, B, or AB) and a Rhesus factor (+ or −).(+ or −). They make up the sample space S:S :

$$S=\{O+,O-,A+,A-,B+,B-,AB+,AB-\}$$

$S = \{O+, O-, A+, A-, B+, B-, AB+, AB-\}$



Janice Haney Carr/CDC

How can we assign probabilities to this sample space? First, the frequencies of these eight blood types differ in different ethnic groups. Within a given ethnic group, we can use the blood types’ frequencies in that group to assign their respective probabilities. The American Red Cross reports that, among Asian Americans, 39% have blood type O+, 1% O−, 27% A+, 0.5% A−, 25% B+, 0.4% B−, 7% AB+, and 0.1% AB −.[3] Because 39% of all Asian Americans have blood type O+, the probability that a randomly chosen Asian American has blood type O+ is 39%, or 0.39. Thus we can use these frequencies to construct the complete probability model for blood types among Asian Americans:

Blood type	O+	O−	A+	A−	B+	B−	AB+	AB−
Probability	0.39	0.01	0.27	0.005	0.25	0.004	0.07	0.001

What if we were interested only in the person’s Rhesus factor? For any randomly selected Asian American, the Rhesus factor can be only positive or negative. Therefore, the sample space for this new question is:

$$S=\{Rh+,Rh-\}$$

$S = \{\mathrm{Rh}+, \mathrm{Rh}-\}$

Based on the known proportions of $\mathrm{Rh}+$ and $\mathrm{Rh}-$ in the Asian American population, the probability model for Rhesus factor is

Rh factor	$\mathrm{Rh}+$	$\mathrm{Rh}-$
Probability	0.98	0.02

EXAMPLE 9.6 Random numbers

Suppose that we want to choose a number at random between 0 and 1. Applet and software random number generators readily perform this task. We might, for example, get the number 0.289351, or the number 0.5816462947. In fact, the sample space is an entire interval of numbers:

$$S = \{\text{all numbers between 0 and 1}\}$$

$S = \{\text{all numbers between 0 and 1}\}$

Let’s call the outcome of the random number generator Y for short. How can we assign probabilities to such events as $\{Y = 1/3\}$ or $\{0.3 \leq Y \leq 0.7\}$? We cannot assign probabilities to each individual value of Y and then add them, because there are infinitely many possible values. Instead, we will need mathematical functions to assign probabilities over the entire sample space.

The sample space $S = \{O+, O-, A+, A-, B+, B-, AB+, AB-\}$ in Example 9.5 is called a **discrete** sample space, whereas the sample space $S = \{\text{all numbers between 0 and 1}\}$ in Example 9.6 is called a **continuous** sample space. A discrete sample space is an itemized list of outcomes, where each outcome has an associated probability. A continuous sample space is an interval of outcomes, and probabilities over the interval are assigned with a mathematical function.

APPLY YOUR KNOWLEDGE

9.5 Family planning. A couple wants to have three children. Assume that the probabilities of a newborn being male or being female are the same and that the gender of one child does not influence the gender of another child.

a. There are eight possible arrangements of girls and boys. What is the sample space for having three children (gender of the first, second, and third child)? All eight arrangements are (approximately) equally likely.

b. The future parents are wondering how many boys they might get if they have three children. Give a probability model (sample space and probabilities of outcomes) for the number of boys. Follow the method of Example 9.5.

9.6 Identify the sample space. A randomly chosen subject arrives for a study of exercise and fitness. Describe a sample space Ss for each of the following situations. Indicate whether the sample space is discrete or continuous. (In some cases, you may have some freedom in your choice of S.s.)

a. You record the gender of the subject.

b. You ask a subject to pedal for 10 minutes on an exercise bicycle and then rate his or her effort on the Rate of Perceived Exertion (RPE) scale. RPE ranges in whole-number steps from 6 (no exertion at all) to 20 (maximal exertion).

c. You measure VO_2, the maximum volume of oxygen consumed per minute during exercise. VO_2 is generally a value between 2.5 and 6 liters per minute.

d. You measure heart rate after the exercise session and record whether or not it exceeds 75 beats per minute.

STATS
IN YOUR
WORLD

WHAT IS THE CHANCE OF RAIN?

We have all checked the weather forecast to see whether it is likely to rain on our weekend plans. A report says that the chance of rain in your city tomorrow is 20%. What does that mean and how is this probability calculated? The National Weather Service keeps a historical database of daily weather conditions such as temperature, pressure, and humidity. The chance of rain on a given day is calculated as the percent of days in the database with similar weather conditions that had rain. So, a 20% chance of rain tomorrow means that it has rained in only 20% of days with similar weather conditions.

PROBABILITY RULES

Examples 9.5 and 9.6 describe simple random phenomena, involving either a discrete sample space or a continuous sample space. The rules of probability are the same in both cases, because they do not depend on whether the sample space is discrete or continuous. These rules follow from the idea of probability as “the long-run proportion of repetitions on which an event occurs.” Here are some important facts that must be true for *any* assignment of probabilities.

1. **Any probability is a number between 0 and 1, inclusively.** Any proportion is a number between 0 and 1, so any probability is also a number between 0 and 1. An event with probability 0 never occurs, and an event with probability 1 occurs on every trial. An event with probability 0.5 occurs in half the trials in the long run.
2. **All possible outcomes of a sample space together must have probability 1.** Because some outcome must occur on every trial, the sum of the probabilities for all possible outcomes of a random phenomenon must be exactly 1.
3. **When two events have no outcomes in common, they can never happen together, which means that their joint probability is zero. Then, the probability that one *or* the other occurs is the sum of their individual probabilities.** If one event occurs in 40% of all trials, a different event occurs in 25% of all trials, and the two can never occur together, then one or the other occurs on 65% of all trials because 40%+25%=65%.$40\% + 25\% = 65\%.$
4. **The probability that an event does not occur is 1 minus the probability that the event does occur.** If an event occurs in (say) 70% of all trials, it fails to occur in the other 30%. The probability that an event occurs and the probability that it does not occur always add to 100%, or 1.

We can use mathematical notation to state Facts 1 to 4 more concisely. Capital letters near the beginning of the alphabet denote events. If AA is any event, we write its probability as P(A).$P(A)$. Here are our probability facts in formal language. As you apply these rules, remember that they are just another form of intuitively true facts about long-run proportions.

PROBABILITY RULES

Rule 1. The probability $P(A)$ of any event A satisfies $0 \le P(A) \le 1$.

Rule 2. If S is the sample space in a probability model, then $P(S) = 1$.

Rule 3. Two events A and B are **disjoint (mutually exclusive)** if their joint probability $P(A \text{ and } B) = 0$, meaning that they have no outcomes in common and so can never occur together. If A and B are disjoint, then

$$P(A \text{ or } B) = P(A) + P(B)$$

$P(A \text{ or } B) = P(A) + P(B)$

This is the **addition rule for disjoint events.**

Rule 4. For any event A,

$$P(A \text{ does not occur}) = 1 - P(A)$$

$P(A \text{ does not occur}) = 1 - P(A)$

where the event “A does not occur” is called the **complement** of event A.

STATS IN YOUR WORLD

AND, OR GO TO COURT

In its case against the maker of green coffee weight-loss supplements for unsubstantiated health claims, the Federal Trade Commission started its deposition with the following definition: *“And” and “or” shall be construed conjunctively or disjunctively as necessary, to make the applicable phrase or sentence inclusive rather than exclusive.* In legal claims, the meaning of these two words must be explained or else parties would argue in court over them. In probability, we have one definition for each term that remains the same regardless of context.

Notice here that the term “or” in $P(A \text{ or } B)$ indicates that at least one event must be satisfied or true (either A alone, or B alone, or both A and B together). In contrast, the term “and” in $P(A \text{ and } B)$ indicates that both events must be satisfied or true at the same time. In English, these two terms can have different interpretations depending on context. This is not the case in probability, where their interpretation is specific and consistent.

The addition rule extends to more than two events that are disjoint in the sense that no two have any outcomes in common. If events A, B, and C are disjoint, the probability that one of these events occurs is $P(A)+P(B)+P(C)$. *When events are not disjoint and therefore can happen together at the same time, you must take care not to count the probability of some events more than once.* This caution will be expressed formally in the general addition rule we will see in Chapter 10 (page 248).

CAUTION

EXAMPLE 9.7 Using the probability rules: Blood types

We already used the addition rule, without calling it by that name, to find the probabilities in Example 9.5. The event “Rh+” contains four disjoint outcomes displayed in the sample space for blood type, so the addition rule (Rule 3) says that its probability is

$$P(\text{Rh+})=P(\text{O+})+P(\text{A+})+P(\text{B+})+P(\text{AB+})=0.39+0.27+0.25+0.07=0.98$$

$$\begin{aligned} P(\text{Rh+}) &= P(\text{O+})+P(\text{A+})+P(\text{B+})+P(\text{AB+}) \\ &= 0.39+0.27+0.25+0.07 \\ &= 0.98 \end{aligned}$$

Check that the two probabilities in Example 9.5, found using the addition rule, are all between 0 and 1 and add to exactly 1. That is, the probability model for the Rhesus factor obeys Rules 1 and 2.

What is the probability that an Asian American’s blood type would not have a positive Rhesus? By Rule 4,

$$P(\text{Rhesus is not positive})=1-P(\text{Rhesus is positive})=1-0.98=0.02$$

$$\begin{aligned} P(\text{Rhesus is not positive}) &= 1-P(\text{Rhesus is positive}) \\ &= 1-0.98=0.02 \end{aligned}$$

APPLY YOUR KNOWLEDGE

9.7 Rabies in Florida. Rabies is a viral disease of mammals transmitted through the bite of a rabid animal. The virus infects the central nervous system, causing encephalopathy and ultimately death. The Florida Department of Health reports the distribution of documented cases of rabies for all of 2016:[4]

Species	Raccoon	Bat	Fox	Other
Probability	0.53	0.22	0.10	?

a. What probability should replace "?" in the distribution?

b. What is the probability that a reported case of rabies is not a raccoon?

c. What is the probability that a reported case of rabies is either a bat or a fox?

Gregory G. Dimijian, M.D./Science Source

9.8 What's your weight group? According to the U.S. Census Bureau, 29% of American males 18 years of age or older are obese, 41% are overweight, 29% have a healthy weight, and 1% are underweight.[5]

a. Does this assignment of probabilities to adult American males satisfy Rules 1 and 2?

b. What percent of adult American males have a weight higher than what is considered healthy?

c. The Census Bureau reports that the percent of adult American females with a weight higher than what is considered healthy is

56%. Does the assignment of probabilities for males and females who are over a healthy weight satisfy Rules 1 and 2?

DISCRETE VERSUS CONTINUOUS PROBABILITY MODELS

Examples 9.5 and 9.7 illustrate how to assign probabilities to events when the sample space is discrete: Assign a probability to every individual outcome, and then add these probabilities to find the probability of any specific event.

DISCRETE PROBABILITY MODEL

A probability model with a sample space made up of a list of individual outcomes[6] is called **discrete.**

To assign probabilities in a discrete model, list the probabilities of all the individual outcomes. These probabilities must be numbers between 0 and 1 and must have sum 1. The probability of any event is the sum of the probabilities of the outcomes making up the event.

EXAMPLE 9.8 Hearing impairment in Dalmatians

Pure dog breeds are often highly inbred, leading to high numbers of congenital defects. A study examined hearing impairment in 5333 Dalmatians.[7] Call the number of ears impaired (deaf) in a randomly chosen Dalmatian Xx for short. The researchers found the following probability model for X:x :

Xx	0	1	2
Probability	0.70	0.22	0.08

Check that the probabilities of the outcomes sum to exactly 1, as they should in a legitimate discrete probability model.

The probability that a randomly chosen Dalmatian has some hearing impairment is the probability that Xx is equal to or greater than 1:

$$P(X\geq 1)=P(X=1)+P(X=2)=0.22+0.08=0.30$$

$$\begin{aligned} P(X \geq 1) &= P(X=1)+P(X=2) \\ &= 0.22+0.08=0.30 \end{aligned}$$

Almost one-third of Dalmatians are deaf in one or both ears. This very high proportion may be explained in part by the fact that breeders cannot detect partial deafness from the dog's behavior. The study suggested giving the dogs a hearing test before considering them for breeding.

Note that the probability that X *is greater than or equal to 1 ("at least 1") is not the same as the probability that* X *is strictly greater than 1.* The latter probability here is

$$P(X>1)=P(X=2)=0.08$$

$P(X>1)=P(X=2)=0.08$

The outcome $X=1$ is included in "greater than or equal to" and is not included in "strictly greater than."

In contrast, Example 9.6 introduced a random phenomenon with a continuous sample space. That sample space is a whole interval, $S=\{\text{all numbers between 0 and 1}\}$, so it contains an infinite number of possible outcomes. We cannot assign probabilities to each individual value in the sample space, but we can use a mathematical function to assign probabilities for any range of values within the sample space. These mathematical functions are called density curves.

DENSITY CURVE

A **density curve** is a curve that

- is always on or above the horizontal axis, and
- has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values on the horizontal axis is the proportion of all observations that fall within that range.

We use a new way of assigning probabilities directly to events in a continuous sample space—as *areas under a density curve*. Density curves are models for continuous distributions. Any density curve has area exactly 1 underneath it, corresponding to total probability 1.

In Chapter 1 we described continuous distributions with histograms. Sometimes the overall pattern of a large number of observations is so regular that we can describe it with a smooth curve.

EXAMPLE 9.9 From histogram to density curve

Figure 9.2 is a histogram of the heights of women 40 to 49 years of age in a large sample survey.[8] Overall, the distribution of heights is quite regular. The histogram is symmetric, and both tails fall off smoothly from a single center peak. There are no large gaps or obvious outliers. The smooth curve drawn over the histogram is a good description of the overall pattern of the data. Our eyes respond to the *areas* of the bars in a histogram; these bar areas represent the percents of the observations. Figure 9.3(a) is a copy of Figure 9.2 with the leftmost bars shaded. The area of the shaded bars represents the women 62 inches tall or less. They make up 31.9% of all women in the sample—this is the cumulative percent, the sum of all bars for heights of 62 inches or less.

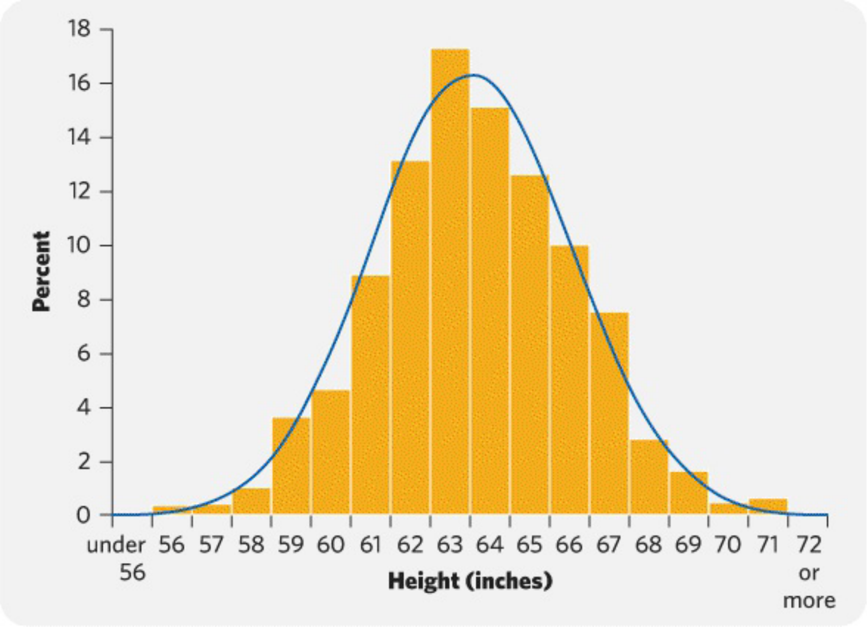


Figure 9.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 9.2 Histogram of the heights in inches of women age 40 to 49 in the United States. The smooth curve shows the overall shape of the distribution.

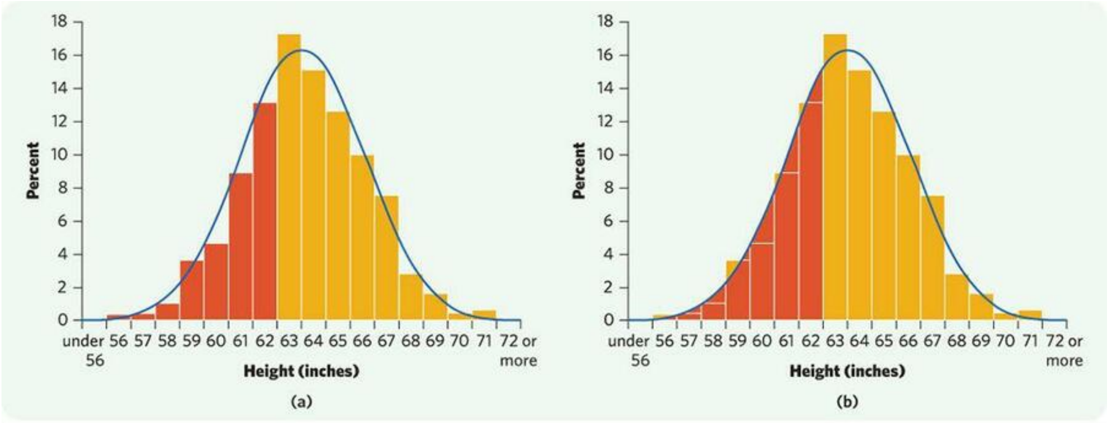


Figure 9.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 9.3 (a) The proportion of women 62 inches tall or less in the sample is 0.319. (b) The proportion of women 62 inches tall or less calculated from the density curve is 0.316. The density curve is a good approximation to the distribution of the data.

Now look at the curve drawn through the bars. In Figure 9.3(b), the area under the curve to the left of 62 inches is shaded. The smooth curve we use to model the histogram distribution is chosen with the specific constraint that *the total area under the curve is exactly 1*. The total area represents 100%—that is, all the observations. We can then interpret areas under the curve as proportions of the observations. The curve is now a *density curve*. The shaded area under the density curve in Figure 9.3(b) represents the proportion of women in their 40s who are 62 inches or shorter. This area is 31.6%, less than half a percentage point away from the actual 31.9%. Areas under the density curve give quite good approximations to the actual distribution of the sampled women.

A density curve is often a good description of the overall pattern of a distribution. Outliers, which are deviations from the overall pattern, are not described by the curve. *Of course, no set of real data is exactly described by a density curve. The curve is a model, an idealized description that is easy to use and accurate enough for practical use*. Conceptually, a density curve is similar to a regression line: We use a least-squares regression line to model an observed linear trend and to make predictions about similar individuals in the population.

How do we use density curves to assign probabilities over continuous intervals? There is a direct relationship between the representation (or proportion) of a given type of individual in a population and the probability that one individual randomly selected from the population will be of that given type. Just as with discrete probabilities, we define probabilities over continuous intervals by the relative frequency of relevant individuals in the population—in this case, all individuals from the population that belong to the desired interval.

EXAMPLE 9.10 From population distribution to probability distribution

In Example 9.9 we used a density curve to estimate that the proportion of women in their 40s who are 62 inches or shorter is 31.6%. This is the area under the density curve for heights of 62 inches or less, as shown in Figure 9.3(b).

Let’s now ask a different question: What is the probability that a randomly chosen woman in her 40s has a height of 62 inches or less? Because the selection is random, this probability depends on the relative frequency of women in the population who are 62 inches tall or less, and that relative frequency is 31.6%. Therefore, the probability that a randomly selected woman in her 40s would measure 62 inches or less is 0.316. This is the area under the density curve for heights 62 inches or less.

CONTINUOUS PROBABILITY MODEL

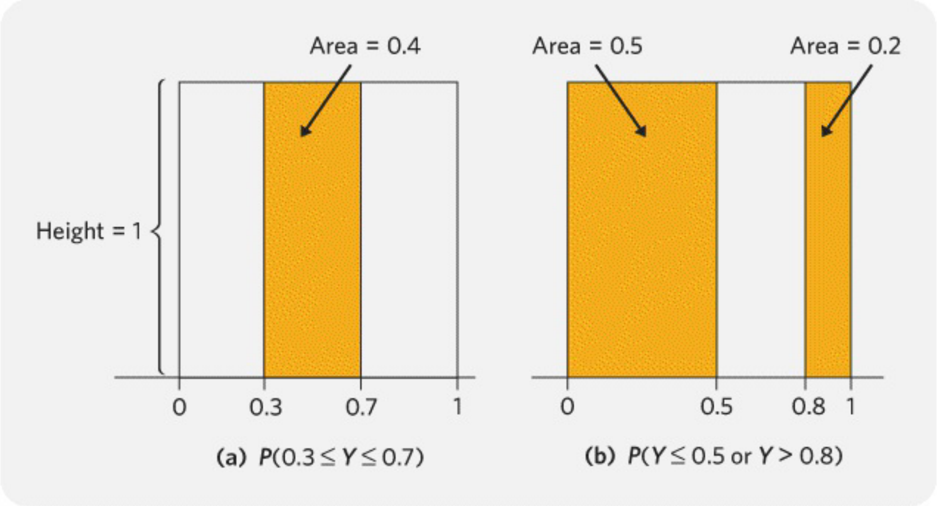
A **continuous probability model** assigns probabilities as areas under a density curve. The area under the curve and above any range of values on the horizontal axis is the probability of an outcome occurring in that range.

Density curves, like distributions, come in many shapes. Example 9.6

described the random outcome of selecting any number between 0 and 1 using a computer random number generator, modeled with a completely different density curve.

EXAMPLE 9.11 Random numbers

The random number generator will spread its output uniformly (evenly) across the entire interval from 0 to 1 as we allow it to generate a long sequence of numbers. The results of many trials are represented by the *uniform density curve* shown in Figure 9.4. This density curve has height 1 over the interval from 0 to 1. The area under the curve is 1, and the probability of any event is the area under the curve and above the event in question.



(a) $P(0.3 \leq Y \leq 0.7)$

(b) $P(Y \leq 0.5 \text{ or } Y > 0.8)$

Figure 9.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 9.4 Probability as area under a density curve. This uniform density curve spreads probability evenly between 0 and 1.

As Figure 9.4(a) illustrates, the probability that the random number generator produces a number between 0.3 and 0.7 is

$$P(0.3 \leq Y \leq 0.7) = 0.4$$

$$P(0.3 \leq Y \leq 0.7) = 0.4$$

because the area under the density curve and above the interval from 0.3 to 0.7 is 0.4. The height of the curve is 1 and the area of a rectangle is the product of height and length, so the probability of any interval of outcomes is just the length of the interval.

Similarly,

P(Y≤0.5)=0.5P(Y>0.8)=0.2P(Y≤0.5 or Y>0.8)=0.7

$$\begin{aligned} P(Y \leq 0.5) &= 0.5 \\ P(Y > 0.8) &= 0.2 \\ P(Y \leq 0.5 \text{ or } Y > 0.8) &= 0.7 \end{aligned}$$

The last event consists of two nonoverlapping intervals, so the total area above the event is found by adding two areas, as illustrated by Figure 9.4(b). This assignment of probabilities obeys all our rules for probability.

The probability model for a continuous random variable assigns probabilities to intervals of outcomes rather than to individual outcomes. In fact, *all continuous probability models assign probability 0 to any individual outcome. Only intervals of values have positive probability.* To see that this is true, consider a specific outcome such as $P(Y = 0.8)$ in Example 9.11. The probability of any interval is the same as its length. The point 0.8 has no length, so its probability is 0.

We can use any density curve to assign probabilities. In Chapter 11 we will examine *Normal curves,* a particularly useful family of density curves in probability and statistics.

APPLY YOUR KNOWLEDGE

9.9 Soda consumption. A survey by Gallup asked a random sample of American adults about their soda consumption.[9] Let's call X the number of glasses of soda consumed on a typical day. Gallup found the following probability model for X:

X	0	1	2	3	4+
Probability	0.52	0.28	0.09	0.04	0.07

Consider the events

A={number of glasses of soda is 1 or greater}B={number of glasses of soda is 2 or less}

$$\begin{aligned} A &= \{\text{number of glasses of soda is 1 or greater}\} \\ B &= \{\text{number of glasses of soda is 2 or less}\} \end{aligned}$$

a. What outcomes make up the event A?A? What is P(A)?$P(A)$?
b. What outcomes make up the event B?B? What is P(B)?$P(B)$?
c. What outcomes make up the event “AA or BB”? What is P(A or B)?$P(A \text{ or } B)$? Why is this probability not equal to P(A)+P(B)?
$P(A) + P(B)$?

age fotostock/SuperStock

9.10 TV viewing among high school students. Some argue that TV viewing promotes obesity because of its association with physical inactivity and high exposure to commercials for foods with high fat and sugar content. In the two examples provided here, decide whether the probability model is discrete or continuous. Explain your reasoning.

a. According to the National Center for Health Statistics, there is a 38% chance that a high school student watches at least 3 hours of television on an average school day.[10] This nationwide survey recorded the amount of time high school students watch TV on an average school day.

b. A large sample survey of school-age children conducted by the Kaiser Family Foundation found a 42% chance that the TV is on most of the time at home.[11] The survey asked whether, at home, the TV was on most of the time, some of the time, a little bit of the time, or never.

9.11 Random numbers. Let X be a random number between 0 and 1 produced by the random number generator described in Example 9.11 and Figure 9.4. Find the following probabilities:

a. $P(X \leq 0.4)$
b. $P(X < 0.4)$
c. $P(0.3 \leq X \leq 0.5)$
d. $P(X < 0.3 \text{ or } X > 0.5)$

RANDOM VARIABLES

Examples 9.8 and 9.11 use a mathematical notation that is often convenient. It is especially helpful when using mathematical functions to describe probability distributions, as we will see in Chapters 11 and 12. In Example 9.8 we let XX stand for the result of choosing a Dalmatian at random and assessing its hearing impairment (number of ears impaired). We know that XX may take a different value if we make another random choice (pick a different Dalmatian). Because its value changes from one random choice to another, we call the hearing impairment XX a *random variable.*

RANDOM VARIABLE

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

The **probability distribution** of a random variable XX tells us what values XX can take and how to assign probabilities to those values.

There are two main types of random variables, corresponding to two types of probability models: *discrete* and *continuous.*

EXAMPLE 9.12 Discrete and continuous random variables

The hearing impairment XX in Example 9.8 is a random variable whose possible values are the whole numbers {0,1,2}.$\{0, 1, 2\}$. The distribution of XX assigns a probability to each of these outcomes. Random variables that have a countable (typically finite) list of possible outcomes are called **discrete.**

Compare this with the value YY obtained with the random number generator in Example 9.11. The values of YY fill the entire interval of numbers between 0 and 1. The probability distribution of YY is given by its density curve, shown in Figure 9.4. Random variables that can take on any value in an interval, with probabilities given as areas under a density

> curve, are called **continuous.**

Be sure to consider closely the true nature of your random variable. Example 9.9 described the process of assigning a density curve to model women’s heights in inches. While most people report their heights in whole numbers of inches, they do not grow one inch at a time. All possible values of height within a realistic interval can actually exist—for example, 65.125 inches. Thus, the random variable height is truly a continuous random variable.

Measures of center and spread apply to random variables as well as to variables in actual sets of observations. The mean and standard deviation of a random variable, in particular, describe important and commonly reported aspects of a probability distribution. Because random variables are mathematical models for all possible outcomes in a sample space S, their mean and standard deviation are parameters rather than statistics. The usual notation for the **mean of a random variable** μ is μ (the Greek letter mu). We write the **standard deviation of a random variable** σ as σ (the Greek letter sigma). We use Greek letters to distinguish these parameters from the statistics $\bar{x}$ and s representing, respectively, the mean and standard deviation computed from actual sample observations.

Computing the mean μ and standard deviation σ of a discrete random variable involves simple enough arithmetic. The process is more complex mathematically for continuous random variables and beyond the scope of an introductory statistics textbook.

COMPUTING THE MEAN μ AND THE STANDARD DEVIATION σ OF A DISCRETE RANDOM VARIABLE

Let X be a discrete random variable with k elements, $x_1, x_2, \dots, x_k$, in its sample space S.

The **mean** of X is

$$\mu = \sum x_i * P(X = x_i)$$

The **variance** of X is[12]

σ2=∑(xi-μ)2*P(X=xi)

$\sigma^2 = \sum (x_i - \mu)^2 * P(X = x_i)$

The **standard deviation** of XX is the square root of its variance: σ=σ2$\sigma = \sqrt{\sigma^2}$

The mean μμ of a random variable XX is also called its **expected value,** because it is the value we expect to get, on average, over many repetitions of the random event. Another notation for the mean μμ is E(X), $E(X)$, which stands for “expected value of X.X.” The standard deviation σσ is a measure of dispersion away from the mean that takes into account both how far each value of XX is from the mean μμ and how likely each value is. Although it is not quite mathematically correct, you can think of the standard deviation σσ of a random variable as roughly the expected deviation from the mean in the long run.

EXAMPLE 9.13 Hearing impairment in Dalmatians

Example 9.8 described the distribution of the random variable XX representing the number of impaired ears (deaf) in a randomly chosen Dalmatian. Researchers found the following probability model for X:X:

XX	0	1	2
Probability	0.70	0.22	0.08

Agency Animal Picture/Getty Images

> The mean of Xx is
>
> μ=∑xi*P(X=xi)μ=0*0.70+1*0.22+2*0.08=0.38
>
> $$\mu = \sum x_i * P(X = x_i)$$
> $$\mu = 0 * 0.70 + 1 * 0.22 + 2 * 0.08 = 0.38$$
>
> The variance of Xx is
>
> σ2=∑(xi-μ)2*P(X=xi)σ2=(0-0.38)2*0.70+(1-0.38)2*0.22+(2-0.38)2*0.08=0.3956
>
> $$\sigma^2 = \sum (x_i - \mu)^2 * P(X = x_i)$$
> $$\sigma^2 = (0 - 0.38)^2 * 0.70 + (1 - 0.38)^2 * 0.22 + (2 - 0.38)^2 * 0.08 = 0.3956$$
>
> The standard deviation of Xx is σ=σ2=0.3956=0.62897.
>
> $\sigma = \sqrt{\sigma^2} = \sqrt{0.3956} = 0.62897.$
>
> In the long run, in very large sets of randomly selected Dalmatian dogs, we expect to find an average of 0.38 impaired ears per dog. The expected deviation from this value is, roughly speaking, 0.63 impaired ears.

Note that the expected value in Example 9.13 is not one of the values that the random variable Xx can take. This is because the expected value is a mean rather than an actual outcome.

Continuous random variables also have means and standard deviations. Actual calculations are different because they require continuous integration, but their interpretation is the same in both situations. We will further discuss the mean and standard deviation of continuous random variable in Chapter 11, in the specific context of Normal distributions.

APPLY YOUR KNOWLEDGE

9.12 Discrete or continuous random variable? Indicate in the following examples whether the random variable Xx is discrete or continuous. Explain your reasoning.

a. Xx is the number of days last week that a randomly chosen child exercised for at least one hour.

b. Xx is the amount of time in hours that a randomly chosen child spends watching television today.

c. X is the number of television sets in a randomly chosen household.

9.13 Physically active high schoolers. The 2011 National Youth Risk Behavior Survey provides insight on the physical activity of high school students in the United States. More than 15,000 high schoolers were asked, “During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?” Physical activity was defined as any activity that increased heart rate. Let X represent the response. The survey results give the following probability model for X:[13]

Days	0	1	2	3	4	5	6	7
Probability	0.15	0.08	0.10	0.11	0.10	0.12	0.07	0.27

a. Verify that this is a legitimate discrete probability model.

b. Describe the event $X<7$ in words. What is $P(X<7)$?

c. Express the event “physically active on at least one day” in terms of X. What is the probability of this event?

d. Obtain the mean μ and standard deviation σ of X.

RISK AND ODDS[i]

Random events can be described in a variety of ways. The most common descriptors of random events in the life sciences are probability, risk, and odds. So far in this chapter, we have introduced the idea of probability and fundamental probability rules. We now briefly describe what risk and odds are and how they relate to the notion of probability. Chapter 20 will further discuss the applications of risk and odds in clinical research when comparing two groups.

Risk means different things in different fields. In statistics, **risk** corresponds to the probability of an undesirable event such as death, disease, or side effects. This term is particularly common in the health sciences, which aim to assess risk and find ways to reduce it. The risk of a given adverse event, like its probability, is defined by the frequency of that adverse event in a population or sample of interest.

Odds are a somewhat less intuitive concept, with a foundation in gambling (although mathematical odds are not to be confused with betting odds, which typically reflect payoffs offered by bookies to winning bets). An **odds** is a ratio of two probabilities, where the numerator represents the probability of an event occurring and the denominator represents the complementary probability of that event not occurring. Therefore, odds can take any positive value, including values greater than 1. The odds of an event can be expressed as the numerical value of the ratio or as a ratio of two integers with no common denominator.

RISK AND ODDS

The **risk** of an *undesirable* outcome of a random phenomenon is the probability of that undesirable outcome.

The **odds** of any outcome of a random phenomenon is the ratio of the probability of that outcome over the probability of that outcome not occurring.

That is, if an outcome AA has probability pp of occurring, then

$$\text{risk (A)=podds (A)=p/(1-p)}$$

$$\begin{aligned} \text{risk } (A) &= p \\ \text{odds } (A) &= p/(1-p) \end{aligned}$$

EXAMPLE 9.14 Blood clots in immobilized patients

Patients immobilized for a substantial amount of time can develop deep vein thrombosis (DVT), a blood clot in a leg or pelvis vein. DVT can have serious adverse health effects and can be difficult to diagnose. On its website, the drug manufacturer Pfizer reports the outcome of a study looking at the effectiveness of the drug Fragmin (dalteparin) in preventing DVT in immobilized patients. Of the 1518 randomly chosen immobilized patients given Fragmin, 42 experienced a complication from DVT (the remaining 1476 patients did not).[14]

The proportion of patients experiencing DVT complications is 42/1518=0.0277, or 2.77%.$42/1518 = 0.0277, \text{ or } 2.77\%.$ We can use this information to compute the risk and odds of experiencing DVT complications for immobilized patients treated with Fragmin:

risk=0.0277, or 2.77%odds=0.0277/(1-0.0277)=42/1476=0.0285

$$\begin{aligned} \text{risk} &= 0.0277, \text{ or } 2.77\% \\ \text{odds} &= 0.0277/(1-0.0277) = 42/1476 = 0.0285 \end{aligned}$$

The odds of experiencing DVT complications among immobilized patients given Fragmin are 42:1476, or about 1:35. That is, for every such patient experiencing a DVT complication, approximately 35 patients do not experience a DVT complication.

The numerical values for the risk and odds in this example are very close, 0.0277 and 0.0285. In general, when the undesirable event is not very frequent, risk and odds give similar numerical values (because the odds denominator, 1-p,$1-p$, is close to 1). In other situations, risk and odds can be very different.

EXAMPLE 9.15 Sickle-cell anemia

Sickle-cell anemia is a serious, inherited blood disease affecting the shape of red blood cells. Individuals with two copies of the gene causing the defect suffer pain from blocked arteries and can have their life shortened from organ damage. Individuals carrying only one copy of the defective gene (“sickle-cell trait”) are generally healthy but may pass the gene to their offspring. An estimated two million Americans carry the sickle-cell trait.

If a couple learns from blood tests that they both carry the sickle-cell trait, the genetic laws of inheritance tell us that there is a 25% chance that they could conceive a child suffering from sickle-cell anemia. That is, the risk of conceiving a child who will have sickle-cell anemia is 0.25, or 25%. The odds of this outcome are

$$\text{odds}=0.25/(1-0.25)=0.333, \text{ or } 1:3$$

$$\text{odds} = 0.25/\,(1 - 0.25) = 0.333, \text{ or } 1:3$$

Janice Haney Carr/CDC

In this example, the risk and the odds of conceiving a child who has sickle-cell anemia have quite different numerical values: 0.25 and 0.33, respectively. Always make sure that you understand how risk and odds relate to probability when you read reports about these concepts.

Lastly, when only the odds are provided, you can compute the risk pp using the following simple formula:

$$p=\text{odds}/(1+\text{odds})$$

$$p = \text{odds} / (1 + \text{odds})$$

In the sickle-cell anemia example, the odds were 1/3. We can find the risk from this information as

p=odds/(1+odds)=131+13=14=0.25

$$p = \text{odds} / (1 + \text{odds}) = \frac{\frac{1}{3}}{1+\frac{1}{3}} = \frac{1}{4} = 0.25$$

APPLY YOUR KNOWLEDGE

9.14 Blood clots in immobilized patients, continued. The study described in Example 9.14 compared patients treated with Fragmin with patients given a placebo in a randomized, double-blind design. Of the 1473 immobilized patients given a placebo, 73 experienced a complication from DVT.

a. Compute the proportion of patients given a placebo who experienced a complication from DVT. What are the risk and the odds of experiencing a complication from DVT when an immobilized patient is given a placebo? How do these values compare?
b. Compare your results with those of Example 9.14. What do you conclude? We will see in Chapter 20 how to formally compare risks or odds for two groups.

9.15 HPV infections in women. Human papillomavirus (HPV) infection is the most common sexually transmitted infection. Certain types of HPV can cause genital warts in both men and women and cervical cancer in women. The U.S. National Health and Nutrition Examination Survey (NHANES) contacted a representative sample of 1921 women between the ages of 14 and 59 years and asked them to provide a self-collected vaginal swab specimen. Of these specimens, 515 tested positive for HPV, indicating a current HPV infection.[15]

a. Give the probability, risk, and odds that a randomly selected American woman between the ages of 14 and 59 years has a current HPV infection.

b. The survey broke down the data by age group:

Age group (years)	14–19	20–24	25–29	30–39	40–49	50–59
Percent HPV positive	24.5	44.8	27.4	27.5	25.2	19.6

Give the risk and the odds of being HPV positive for women in each age group. Which age group has the highest risk, and has the highest odds, of testing positive for HPV?

CHAPTER 9 SUMMARY

- A **random phenomenon** has outcomes that we cannot predict, but nonetheless has a regular distribution of outcomes in very many repetitions.
- A **probability model** for a random phenomenon consists of a sample space SS and an assignment of probabilities P.P.
- The **sample space** SS is the set of all possible outcomes of the random phenomenon. Sets of outcomes are called **events.** PP assigns a number P(A)$P(A)$ to an event AA as its probability.
- In a frequentist interpretation, the **probability** of an event is the proportion of times the event occurs in many repeated trials of a random phenomenon. This probability can be derived from observed frequencies in a population or from scientific theory. An approximate probability may be derived from observed frequencies in a very large random sample.
- **Personal probabilities** represent an individual's judgment of how likely an outcome is, rather than its frequency in a long run of trials.
- Any assignment of probability must obey the rules that state the basic properties of probability:

1. 0≤P(A)≤1$0 \leq P(A) \leq 1$ for any event A.A.
2. P(S)=1.$P(S) = 1$.
3. **Addition rule:** Events AA and BB are **disjoint** (or mutually exclusive) if P(A and B)=0,$P(A \text{ and } B) = 0$, that is, they have no outcomes in common. If AA and BB are disjoint, then P(A or B)=P(A)+P(B).$P(A \text{ or } B) = P(A) + P(B)$.
4. For any event A,A, P(A does not occur)=1-P(A).
   $P(A \text{ does not occur}) = 1 - P(A)$.

- When a sample space SS contains a finite number of possible values, a **discrete probability model** assigns each of these values a probability between 0 and 1 such that the sum of all the probabilities is exactly 1.

The probability of any event is the sum of the probabilities of all the values that make up the event.

- A sample space can contain all values in some interval of numbers. A **continuous probability model** assigns probabilities as areas under a density curve. A **density curve** has total area 1 underneath it. The probability of any event is the area under the curve and above the values on the horizontal axis that make up the event.
- A density curve is an idealized description of the overall pattern of a distribution that smooths out the irregularities in the actual data. We write the **mean of a density curve** as μμ and the **standard deviation of a density curve** as σσ to distinguish them from the mean x¯ $\bar{x}$ and standard deviation ss of actual sample data.
- A **random variable** is a variable taking numerical values determined by the outcome of a random phenomenon. The **probability distribution** of a random variable, X,X, tells us what the possible values of XX are and how probabilities are assigned to those values. The **mean** μμ (also called the **expected value**) and the **standard deviation** σσ are two important summary values of the probability distribution of a random variable.
- A random variable XX and its distribution can be **discrete** or **continuous.** A **discrete random variable** has distinct values that can be listed. Its distribution gives the probability of each value. A **continuous random variable** takes all values in some numerical interval. A density curve describes the probability distribution of a continuous random variable.
- A probability can also be described as a **risk** when it applies to an undesirable outcome. An **odds** is the ratio of the probability of an outcome occurring over the probability of that outcome not occurring.

THIS CHAPTER IN CONTEXT

This chapter begins our study of probability. The important fact is that random phenomena are unpredictable in the short run but have a regular and predictable behavior in the long run. Probability rules and probability models provide the tools for describing and predicting the long-run behavior of random phenomena.

In Chapter 10 we will explore probability rules in greater depth, including the use of conditional probabilities for medical screening. In Chapters 11 and 12 we will study some important random variables with a wide range of practical and theoretical applications: Normal, binomial, and Poisson.

Probability helps us understand why we can trust random samples and randomized comparative experiments, the subjects of Chapters 6 and 7. It is the key to generalizing what we learn from data produced by random samples and randomized comparative experiments to some wider universe or population. Chapter 13 will focus specifically on the probability properties linking random samples and populations. This will be the founding principle for statistical inference, which we will address throughout the textbook starting with Chapter 14.

CHECK YOUR SKILLS

9.16 You read that in native Hawaiians, the probability of having blood type AB is 1/100 (1 in 100). What does this mean?

a. If you pick 100 Hawaiians randomly, the fraction of them having blood type AB will be very close to 1/100.
b. If you pick 100 Hawaiians randomly, exactly 1 of them will have blood type AB.
c. If you pick 10,000 Hawaiians randomly, exactly 100 of them will have blood type AB.

9.17 A cat is about to have 6 kittens. The sample space for counting the number of female kittens she has is

a. S$\mathcal{S}$ = any number between 0 and 1.
b. S$\mathcal{S}$ = whole numbers 0 to 6.
c. S$\mathcal{S}$ = all sequences of 6 males or females by order of birth, such as FMMFFF.

9.18 Here is the probability model for the blood type of a randomly chosen person of Hispanic ethnicity in the United States, according to the Red Cross:

Blood type	O	A	B	AB
Probability	0.57	0.31	0.10	?

The probability that a randomly chosen American of Hispanic ethnicity has type AB blood

a. can be any number between 0 and 1.
b. is 0.02.
c. is 0.2.

9.19 In a table of random digits such as Table A, each digit is equally likely to be any of 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9. What is the probability that a digit in the table is 7 or greater?

a. 7/10
b. 4/10
c. 3/10

A study of freely forming groups in bars throughout Europe examined the number of individuals found in groups who were laughing together. Let X be the number of individuals in laughing groups.[16] *Here is the probability model for X. Exercises 9.20 to 9.23 use this information.*

Number of individuals XX	2	3	4	5	6
Probability	0.51	0.34	0.10	0.04	0.01

9.20 This probability model is

a. continuous.
b. discrete.
c. discretely continuous.

9.21 What is the probability $P(X \leq 3)$?$P(X \leq 3)$?

a. 0.34
b. 0.51
c. 0.85

9.22 What is the mean $\mu$$\mu$ of X?X?

a. 0.2
b. 2.7
c. 4.0

9.23 What is the standard deviation $\sigma$$\sigma$ of X?X?

a. 0.77
b. 0.88
c. 1.41

The Mental Health Surveillance Study uses trained mental health clinicians to obtain a detailed mental health assessment of a very large random sample of adults. The study found that 17.9% of adults had received a diagnosis of mental illness within the 12 months preceding the interview.[17] *Exercises 9.24 and 9.25 use this information.*

9.24 For adults in the United States, what is the risk of receiving a diagnosis of mental illness in a given year?

a. 0.147
b. 0.179
c. 0.218

9.25 For adults in the United States, what are the odds of receiving a diagnosis of mental illness in a given year?

a. 0.147
b. 0.179
c. 0.218

CHAPTER 9 EXERCISES

9.26 Sample spaces. In each of the following situations, describe a sample space SS for the random phenomenon. In some cases, you have some freedom in your choice of S.S.

a. A seed is planted in the ground. It either germinates or fails to grow.
b. A patient with a usually fatal form of cancer is given a new treatment. The response variable is the length of time that the patient lives after treatment.
c. A nutrition researcher feeds a new diet to a young male white rat. The response variable is the weight (in grams) that the rat gains in eight weeks.
d. Buy a hot dog and record how many calories it has.

9.27 Manatee deaths. Manatees are an endangered species of herbivorous, aquatic mammals found primarily in the rivers and estuaries of Florida. As part of its conservation efforts, the Florida Fish and Wildlife Commission records the cause of death for every recovered manatee carcass in its waters. Of the 392 recorded manatee deaths in 2012, 68 were perinatal and 81 were caused by collision with a watercraft.[18]

a. What is the probability that the death of a randomly selected manatee was due to collision with a watercraft?
b. What is the probability that the death was not due to collision with a watercraft?
c. What is the probability that the cause of death was due to perinatal problems or collision with a watercraft? What is the probability that it was due to some other cause?

9.28 Probability models? In each of the following situations, state whether the given assignment of probabilities to individual outcomes is legitimate—that is, whether it satisfies the rules of probability. If not, give specific reasons for your answer.

a. Among Navajo Native Americans, the distribution of blood type is[19] P(A)=0.27,P(B)=0.00,P(AB)=0.00,P(O)=0.73. $P(A) = 0.27, P(B) = 0.00, P(\text{AB}) = 0.00, P(O) = 0.73.$

b. Choose a cat at random from an animal shelter and record sex and hair type:

   P(female short-haired)=0.56,$P(\text{female short-haired}) = 0.56,$

   P(female long-haired)=0.24,$P(\text{female long-haired}) = 0.24,$

   P(male short-haired)=0.44,P(male long-haired)=0.16. $P(\text{male short-haired}) = 0.44, P(\text{male long-haired}) = 0.16.$

c. A hospital tests every newborn for phenylketonuria (PKU). PKU is one of the most common genetic disorders, easily treated but resulting in brain damage if undiagnosed. The incidence of PKU varies all over the world, as follows:[20] P(United States)=1/10,000,$P(\text{United States}) = 1/10,000,$ P(Ireland)=1/4,500,$P(\text{Ireland}) = 1/4,500,$ P(Finland)=1/100,000, $P(\text{Finland}) = 1/100,000,$ P(rest of the world)=1/15,000. $P(\text{rest of the world}) = 1/15,000.$

9.29 Causes of death among young people. Government data assign a single cause for each death that occurs in the United States. The data show that, among persons aged 15 to 24 years, the probability is 0.41 that a randomly chosen death was an accident, 0.16 that it was a homicide, and 0.15 that it was a suicide.

a. What is the sample space for the probability model of major causes of death in this age group based on the information you are given?

b. What is the probability that a death was either an accident, a homicide, or a suicide? What is the probability that the death was due to some other cause?

9.30 Alcoholic beverages. A 2016 Gallup survey asked a random sample of 1023 American adults if they drink alcoholic beverages and, if they do, which type they consume most often. The survey found that, among those who do drink alcohol, 43% favor beer, 32% favor wine, 20% favor liquor, and the remaining expressed no particular preference.[21] Use these sample proportions to assign approximate probability values for the corresponding events.

a. Give the sample space SS and the probability model for the random event “type of alcoholic beverage favored.”

b. Obtain the probability that a randomly selected American adult who drinks alcohol does not favor beer.

c. Obtain the probability that a randomly selected American adult who drinks alcohol favors either beer or wine.

9.31 Hepatitis C. The CDC provides the breakdown of the sources of infection leading to hepatitis C in Americans. Here are the probabilities of each infection source for a randomly chosen individual with hepatitis C:[22]

Source of infection	**Probability**
Intravenous drug use	0.60
Unprotected sex	0.15
Transfusion (before screening)	0.10
Unknown or other	0.11
Occupational	?

a. What is the probability that a person with hepatitis C was infected in the course of his or her professional occupation?

b. What is the probability that a person with hepatitis C was infected through a known risky behavior (intravenous drug use or unprotected sex)?

9.32 Birth order. The National Vital Statistics database reports all births in 2011 by the reported birth order of the child (first child of the mother, second, and so on). Here is how the 3,923,381 documented births break down:

Birth order	**Percent**
First child	40.2
Second child	31.6
Third child	16.5
Fourth child or later	11.7

a. What is the probability that a randomly selected baby born in 2011 wasn't a woman's first child?

b. What is the probability that a randomly selected woman giving birth in 2011 had given birth to exactly one child before? Think carefully about the birth order of her latest baby.

c. What is the probability that a randomly selected woman giving birth in 2011 had given birth to fewer than two children before?

9.33 Treating diabetes. Insulin injections and oral medication are two key options in diabetes treatment. The 2014 National Diabetes Statistics Report provides the breakdown of diabetes treatments among American adults with diagnosed diabetes, as shown in the following table:[23]

Diabetes treatment	Percent
Insulin only	14.0
Both insulin and oral medication	14.7
Oral medication only	56.9
Neither insulin nor oral medication	14.4

a. Explain why this is a proper probability model. Is the sample space discrete or continuous?

b. What is the probability that a randomly selected American adult diagnosed with diabetes is treated with insulin, either alone or combined with oral medication?

c. What is the probability that a randomly selected American adult diagnosed with diabetes is not treated with insulin at all?

9.34 Osteoporosis. Choose a postmenopausal woman at random. The probability is 0.40 that the woman chosen has osteopenia (low bone density) and 0.07 that she has osteoporosis (pathologically low bone density). Otherwise, her bone density is considered healthy.[24]

a. What is the probability that a randomly chosen postmenopausal woman has a healthy bone density?

b. What is the probability that a randomly chosen postmenopausal woman has some problem with low bone density (osteopenia or osteoporosis)?

9.35 Dentist appointment. A large governmental survey of dental care among children 2 to 17 years old found the following distribution of times since the last visit to the dentist:[25]

Time since last visit	Probability
6 months or less	0.57
More than 6 months but no more than 1 year	0.18
More than 1 year but no more than 2 years	0.08
More than 2 years but no more than 5 years	0.03
More than 5 years	?

a. What is the probability that a child has not seen a dentist in more than 5 years?

b. What is the probability that a randomly chosen child has not seen a dentist within the last 6 months?

c. What is the probability that the child has seen a dentist within the last year?

9.36 Overweight? Choose an American adult at random. Define two events:

A=the person chosen is obese$A = \text{the person chosen is obese}$
B=the person chosen is overweight but not obese
$B = \text{the person chosen is overweight but not obese}$

According to the 2011 National Health Interview Survey, P(A)=0.28 $P(A) = 0.28$ and P(B)=0.35.$P(B) = 0.35.$

a. Explain why events AA and BB are disjoint.

b. Say in plain language what the event “A or B$A \text{ or } B$” is. What is P(A or B)?$P(A \text{ or } B)?$

c. If CC is the event that the person chosen has normal weight or less, what is P(C)?$P(C)?$

9.37 Tracking gonorrhea. To track epidemics, the CDC requires physicians to report all cases of important transmissible diseases. In 2014, for instance, a total of 350,062 cases of the sexually transmitted disease gonorrhea were officially reported. Here is how they break down by the age group of the diagnosed patient.[26]

Age group (years)	**Percent of gonorrhea cases**
0−14	1
15−19	19
20−24	33
25−29	20
30−39	17
40 or older	10

a. Explain why this is a proper probability model. Is the sample space discrete or continuous?
b. What is the probability that a randomly selected gonorrhea case that year was a patient in his or her 20s (age 20 to 29)?

9.38 Eye color in hawks. An observational study examined iris coloration in a population of North American Cooper's hawks. Iris colors range from yellow to red and are numerically coded to reflect pigmentation intensity.[27] Among female hawks, the coded eye color XX has the following distribution:

Eye color	Yellow	Light orange	Orange	Dark orange	Red
XX	1	2	3	4	5
Probability	0.05	0.29	0.47	0.18	0.01

a. As defined here, is the random variable XX discrete or continuous? Why?
b. Write the event "eye color at least 4" in terms of X.$X.$ What is the probability of this event?
c. Describe the event X≤2$X \le 2$ in words. What is its probability? What is the probability that X<2?$X < 2?$

9.39 Eye color in hawks, continued. The study in the Exercise 9.38 also gave the distribution of eye color XX among male hawks:

Eye color	Yellow	Light orange	Orange	Dark orange	Red

X	1	2	3	4	5
Probability	0.00	0.11	0.40	0.32	0.17

a. Write the event “eye color less than 2 or more than 4” in terms of X. What is the probability of this event?

b. Describe the event $2<X\leq 4$ in words. What is its probability?

9.40 Mental health in America. In the Mental Health Surveillance Study, clinicians interviewed a very large random sample of American adults so as to diagnose the presence of mental disorders in the prior 12 months. There are many types of mental disorders, and a person can suffer from more than one. Let X represent the number of mental disorders a person has experienced in the past year. Based on this study, we can assign the following probability distribution to X in the American adult population:[28]

X	0	1	2	3	4 or more
$P(X)$	0.775	0.149	0.041	0.022	0.013

a. Explain why this is a proper probability model. Is the sample space discrete or continuous?

b. Obtain the following probabilities: $P(X<2), P(X\leq 2), P(X>2)$.
$P(X<2), P(X\leq 2), P(X>2)$.

c. Obtain the probability that a randomly selected American adult suffers from at least one mental disorder. From either one or two mental disorders.

9.41 Social media use. A 2015 Pew Research Center study examined the use of 5 social media sites (Facebook, Twitter, Instagram, Pinterest, LinkedIn) among American adult internet users. Let X represent the number of these social media sites used by any randomly selected American adult internet user. Based on the study findings, we may use the following probability distribution for X:

X	0	1	2	3	4	5
$P(X)$	0.20	0.28	0.24	0.16	0.08	0.04

a. Obtain the probability that an American online adult uses at most one of these sites. Obtain the probability $P(2 \le X < 4)$.$P(2 \le X < 4)$.
b. Obtain the mean $\mu$$\mu$ of $X$$X$ and interpret this value in context.
c. Obtain the standard deviation $\sigma$$\sigma$ of X.X. Give an approximate interpretation for this value.

9.42 Understanding density curves. Recall that it is the areas under a density curve, not the height of the curve, that give proportions in a distribution. To illustrate how this works, sketch a density curve that has its peak at 0 on the horizontal axis but has greater area within 0.25 on either side of 1 than within 0.25 on either side of 0.

9.43 Random numbers. Many random number generators allow users to specify the range of the random numbers to be produced. Suppose that you specify that the random number $Y$$Y$ can take any value between 0 and 2. Then the density curve of the outcomes has constant height between 0 and 2, and height 0 elsewhere.

a. Is the random variable $Y$$Y$ discrete or continuous? Why?
b. What is the height of the density curve between 0 and 2? Draw a graph of the density curve.
c. Use your graph from part b and the fact that probability is area under the curve to find $P(Y \le 1)$.$P(Y \le 1)$.

9.44 More random numbers. Find these probabilities as areas under the density curve you sketched in Exercise 9.43.

a. $P(0.5 < Y < 1.3)$$P(0.5 < Y < 1.3)$
b. $P(Y \ge 0.8)$$P(Y \ge 0.8)$

9.45 Anorexia. Exercise 1.8 (page 19) introduced a Canadian study of the age at onset of anorexia nervosa—an eating disorder—for 691 diagnosed adolescent girls. Figure 9.5 shows the density curve (solid line) summarizing the histogram of the data (dotted lines). The density curve has been chopped into three parts shown in different shades.

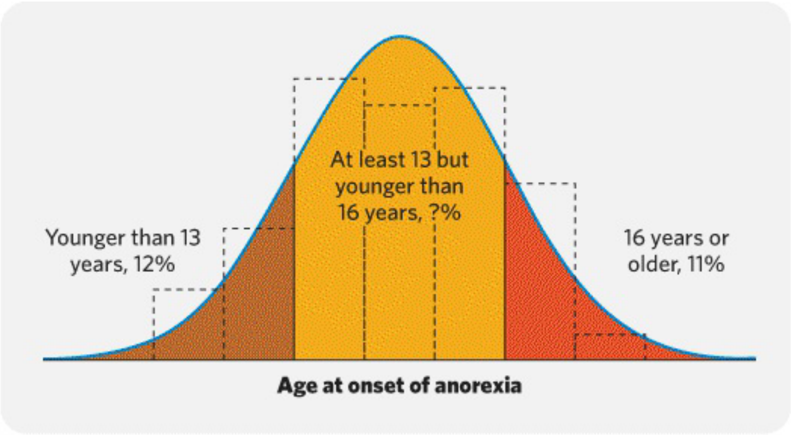


Figure 9.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 9.5 Distribution of age of onset of anorexia nervosa, pictured by both a histogram and a density curve.

a. If we let X represent the age at onset of anorexia, describe the event $13 \le X < 16$ in words. What is its probability?
b. What is the probability that the age at onset for a randomly chosen adolescent girl is sometime after the 13th birthday? Write this event in terms of X.

9.46 Osteoporosis, continued (optional). Go back to Exercise 9.34.

a. What are the risk and the odds of osteopenia in postmenopausal women?
b. What are the risk and the odds of osteoporosis in postmenopausal women?

9.47 Breast cancer (optional). The National Cancer Institute (NCI) compiles U.S. epidemiology data for a number of different cancers (through the Surveillance, Epidemiology, and End Results Program [SEER]). Among women, breast cancer is one of the most common and deadliest types of cancer.

a. Based on SEER data, the NCI estimates that a woman in her 30s has a 0.43% chance of being diagnosed with breast cancer.[29] Compute the risk and the odds of being diagnosed with breast cancer for a randomly chosen woman in her 30s.

b. The risk of breast cancer increases with age. The NCI estimates that a woman in her 60s has a 3.65% chance of being diagnosed with breast cancer. Compute the risk and the odds of being diagnosed with breast cancer for a randomly chosen woman in her 60s.

c. The NCI also reports a lifetime risk of breast cancer—that is, the probability that a woman born in the United States today will develop breast cancer at some time in her life, based on current rates of breast cancer. The most recent estimate of lifetime risk is 12.7%. Give the probability, risk, and odds of developing breast cancer over a lifetime for a randomly chosen woman born in the United States today.

d. Compare the lifetime values you got in part c with the values by decade you found in parts a and b. Which way of communicating risk would you say is most informative (lifetime or by decade)?

9.48 Hospital births. The proportion of baby boys born is approximately 0.5 in the general population. Use the *Probability* applet or software to simulate 100 consecutive births at a local hospital, with probability 0.5 of a boy in each birth. (In most software, the key phrase to look for is “Bernoulli trials.” This is the technical term for independent trials with yes/no outcomes. Our outcomes here are “boy” and “girl.”)

a. What percent of the 100 births are boys?

b. Examine the sequence of boys and girls. How long was the longest run of boys? Of girls? (Sequences of random outcomes often show runs longer than our intuition thinks likely.)

9.49 Simulating a survey. The culinary herb cilantro is very polarizing: Some people love it, but others hate it. A large survey of American adults of European ancestry found that 26% disliked the taste of cilantro.[30] Suppose that this holds exactly true for all American adults of European ancestry. Choosing such an individual at random then has probability 0.26 of getting one who dislikes cilantro. Use the *Probability* applet or your statistical software to simulate choosing many people at random. (In most software, the key phrase to look for is “Bernoulli trials.” This is the technical term for independent trials with yes/no outcomes. Our outcomes here are “dislikes cilantro” and “does not dislike cilantro.”)

a. Simulate drawing 25 people, then 100 people, then 400 people (click “Toss” twice with 200 selected). What proportion, in each simulated sample, dislike cilantro? We expect (but because of chance variation we can’t be sure) that the proportion will be closer to 0.26 in larger samples.
b. Simulate drawing a sample of 25 people 10 times (click “Reset” after each simulation) and record the percents in each simulated sample who dislike cilantro. Then simulate drawing a sample of 400 people 10 times and again record the 10 percents. Which set of 10 results is less variable (the set of samples of 25 or the set of samples of 400)? We expect the results of 400 drawings to be more predictable (less variable) than the results of 25 drawings. These simulations illustrate the concept of “long-run regularity” in random events.

CHAPTER 10 Independence and Conditional Probabilities[i]

Taylor S. Kennedy/Getty Images

IN THIS CHAPTER WE COVER...

- Relationships among several events
- Conditional probability
- General probability rules
- Tree diagrams
- Bayes's theorem
- *Discussion: Making sense of conditional probabilities in diagnostic tests*

The mathematics of probability can provide models to describe the genetic makeup of populations, the spread of epidemics, and risk factors for a given disorder. Although we are interested in probability because of its usefulness in statistics, the mathematics of chance is important in many fields of study. Our study of probability in Chapter 9 concentrated on basic ideas and facts needed to describe one random phenomenon at a time. Now we look at further properties of probability that will allow us to study how different random phenomena relate to one another.

RELATIONSHIPS AMONG SEVERAL EVENTS

In Chapter 9 we described the probability that *one or the other* of two events (A,B)(A, B) occurs in the special situation when these two events cannot occur together: when the two events are *disjoint*, or *mutually exclusive*, and therefore P(A and B)=0.$P(A \text{ and } B) = 0.$

You may find it helpful to draw a picture to display relationships among several events. A picture like Figure 10.1, which shows the sample space SS as a rectangular area and events as areas within S,S, is called a **Venn diagram.** In a Venn diagram, the areas representing events are supposed to be scaled to visually represent the probability of these events. The events AA and BB in Figure 10.1 are disjoint because they do not overlap. Because they do not overlap, there is no area for the event {A and B},$\{A \text{ and } B\}$, which corresponds to the fact that P(A and B)=0.$P(A \text{ and } B) = 0.$

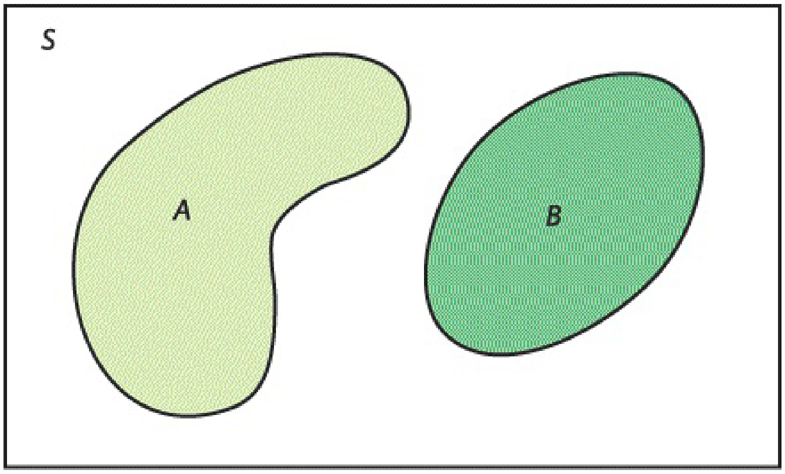


Figure 10.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 10.1 Venn diagram showing disjoint events A and B.

$A \text{ and } B.$ Their joint probability, P(A and B),

$P(A \text{ and } B),$ equals zero.

The Venn diagram in Figure 10.2 illustrates a different situation, in which two events are not disjoint. The event {A and B}$\{A \text{ and } B\}$ appears as the overlapping area that is common to both AA and B,$B,$ indicating a nonzero probability.

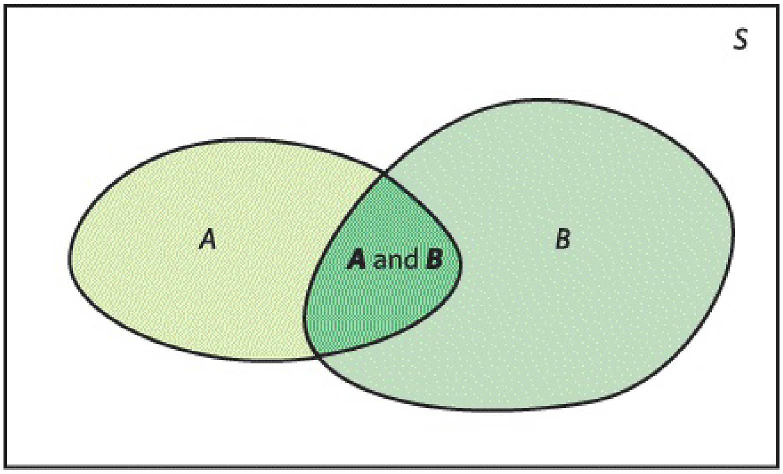


Figure 10.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 10.2 Venn diagram showing events A and B $A \text{ and } B$ that are not disjoint. The event {A and B} $\{A \text{ and } B\}$ consists of outcomes common to A and B.

A and B.

Suppose that you want to study the next two single births at a local hospital. You are counting girls, so two events of interest are

A=first baby is a girlB=second baby is a girl

$$\begin{aligned}A &= \text{first baby is a girl}\\ B &= \text{second baby is a girl}\end{aligned}$$

The events A A and B B are not disjoint (they are not mutually exclusive). They occur together whenever the next two single births at the hospital are girls. We want to find the probability of the event {A and B} $\{A \text{ and } B\}$ that *both* babies are girls. A Venn diagram like that of Figure 10.2 is a useful draft representation. Alternatively, you may prefer displaying all possible combinations of outcomes in a two-way table:

	Second baby is a girl {B}$\{B\}$	**Second baby is not a girl {not B}$\{\text{not } B\}$**
First baby is a girl {A}$\{A\}$	P(A and B)$P(A \text{ and } B)$	P(A and not B)$P(A \text{ and not } B)$
First baby is not a girl {not A}$\{\text{not } A\}$	P(not A and B)$P(\text{not } A \text{ and } B)$	P(not A and not B)$P(\text{not } A \text{ and not } B)$

What we are looking for is P(A and B).$P(A \text{ and } B)$. Using the theoretical, simple model based on the genetic laws described at the beginning of Chapter 9, we assign probability 1/2 to the outcome “the baby is a girl” when any baby is born. So

P(A)=0.5P(B)=0.5

$$\begin{aligned}P(A) &= 0.5\\ P(B) &= 0.5\end{aligned}$$

Common sense says that P(A and B)$P(A \text{ and } B)$ should be 1/4, because the first baby will be a girl half the time; and when the first baby is a girl, the second baby will be a girl half the time. In the long run, then, both babies will be girls in 1/2×1/2=1/4$1/2 \times 1/2 = 1/4$ of all births (half of half).

STATS IN YOUR WORLD

WE WANT A BOY

Misunderstanding independence can be disastrous. “Dear Abby” once published a letter from a mother of eight girls. She and her husband had planned a family of four children. When all four were girls, they kept trying. After seven girls, her doctor assured her that “the law of averages was in our favor 100 to 1.” Unfortunately, having children is like tossing coins. Eight girls in a row is highly unlikely, but once you have seven girls, it is not at all unlikely that the next child will be a girl—and it was.

This reasoning assumes that the second baby still has probability 1/2 of being a girl after the first one born was a girl. This is true—the birth outcome of one couple is not influenced by the birth outcome of another couple. We say that the events “girl on the first birth” and “girl on the second birth” are independent. **Independence** means that the outcome of the first event cannot influence the outcome of the second event.

EXAMPLE 10.1 Independent or not?

Because a coin has no memory and most coin tossers cannot influence the fall of the coin, it is safe to assume that successive coin tosses are independent. For a balanced coin, this means that after we see the outcome of the first toss, we still assign probability 1/2 to heads on the second toss.

In contrast, the colors of successive cards dealt from the same deck are not independent. A standard 52-card deck contains 26 red cards and 26 black cards. For the first card dealt from a shuffled deck, the probability of a red card is 26/52=0.50$26/52 = 0.50$ (equally likely outcomes). Once we see that the first card is red, we know that there are only 25 reds among the remaining 51 cards. The probability that the second card is red is therefore only 25/51=0.49.$25/51 = 0.49$. Knowing the outcome of the first card dealt changes the probability for the second.

If a nurse measures your height twice, it is reasonable to assume that the two results are independent observations. Each event records your

> actual height plus a measurement error (positive or negative), and the size of the error in the first result does not influence the instrument that makes the second reading. But if you take an IQ test twice in succession, the two test scores are not independent, because the learning that occurs on the first attempt influences your second attempt.

MULTIPLICATION RULE FOR INDEPENDENT EVENTS

Two events A and B are **independent** if knowing that one occurs does not change the probability that the other occurs.

If A and B are independent,

$$P(A \text{ and } B) = P(A)P(B)$$

$P(A \text{ and } B) = P(A)P(B)$

Conversely, if this condition is not satisfied, then events A and B are **dependent.**

The multiplication rule also extends to collections of more than two events, provided that all are independent. Independence of events $A, B,$ and C means that no information about any one or any two can change the probability of the remaining events.

If two events A and B are independent, the event that A does not occur is also independent of B, and so on. For example, 45% of all Americans have blood type O.[1] If the Census Bureau interviews two individuals chosen independently, the probability that the first is type O and the second is not type O is $(0.45)(0.55) = 0.2475.$

Independence is often assumed in setting up a probability model when the events we are describing seem to have no logical connection. *However, assuming independence by default is a common mistake and results in faulty computations when the assumption is incorrect.* Do not assume that two events are independent unless you are told that they are or unless the events have no logical connection (for example, when two individuals are independently selected from a large population).

EXAMPLE 10.2 Rapid HIV testing

STATE: Many people who come to clinics to be tested for human immunodeficiency virus (HIV), the virus that causes acquired immunodeficiency syndrome (AIDS), do not return to learn the test results. Clinics now use “rapid HIV tests” that give a result while the client waits. In a clinic in Malawi, for example, use of rapid tests increased the percent of clients who learned their test results from 69% to 99.7%.

The trade-off for fast results is that rapid tests are less accurate than the slower laboratory tests. When applied to people who have no HIV antibodies, one rapid test has a probability of approximately 0.004 of producing a false-positive (that is, of falsely indicating that antibodies are present).[2] If a clinic tests 200 people who are free of HIV antibodies, what is the chance that at least one false-positive will occur?

PLAN: It is reasonable to assume that the test results for different individuals are independent. We have 200 independent events, each with probability 0.004. What is the probability that at least one of these events occurs?

SOLVE: “At least one” combines many outcomes. It is much easier to use the fact that

P(at least one positive)=1-P(no positives)

$$P(\text{at least one positive}) = 1 - P(\text{no positives})$$

and find P(no positives)$P(\text{no positives})$ first.

The probability of a negative result for any one person who has no HIV antibodies is 1-0.004=0.996.$1 - 0.004 = 0.996$. To find the probability that all 200 such persons tested have negative results, use the multiplication rule:

P(no positives)=P(all 200 negative)=(0.996)(0.996)···
(0.996)=0.996200=0.4486

$$\begin{aligned} P(\text{no positives}) &= P(\text{all 200 negative}) \\ &= (0.996)(0.996)\cdots(0.996) \\ &= 0.996^{200} = 0.4486 \end{aligned}$$

The probability we want is therefore

> $$P(\text{at least one positive}) = 1 - 0.4486 = 0.5514$$
>
> **CONCLUDE:** The probability is greater than 1/2 that at least one of the 200 people will test positive for HIV even though none of them has the virus.

The multiplication rule $P(A \text{ and } B) = P(A)P(B)$ *holds if* A *and* B *are independent, but not otherwise. The addition rule* $P(A \text{ or } B) = P(A) + P(B)$, *discussed in Chapter 9, holds if A and B are disjoint, but not otherwise.* Resist the temptation to use these simple rules when the circumstances that justify them are not present. *You must also be careful not to confuse disjointness and independence.* If A and B are disjoint, then the fact that A occurs tells us that B cannot occur—look again at Figure 10.1. Thus, disjoint events are not independent; disjoint events are dependent. Unlike disjointness, we cannot “see” independence in a Venn diagram, because it involves the probabilities of the events rather than just the outcomes that make up the events.

APPLY YOUR KNOWLEDGE

10.1 Spinal cord injuries. The state of Florida reports that 75% of all patients first diagnosed with a spinal cord injury are males.[3]

a. What is the probability that the next 2 patients diagnosed with a spinal cord injury in Florida will be males?

b. Of the next 5 patients diagnosed with a spinal cord injury in Florida, what is the probability that at least 1 will be female? What is the probability that at least 1 will be male?

10.2 Chronic diseases. A Gallup-Healthways survey of more than 350,000 American adults in 2011 found that 30.0% had been told by a physician they have high blood pressure and that 26.2% had been told by a physician that they have a high cholesterol level. Nonetheless, we can’t conclude that because $(0.300)(0.262)=0.0786$,$(0.300)(0.262) = 0.0786$, approximately 7.9% of American adults have been told by a physician that they have high blood pressure and that they have high cholesterol. Why not?

10.3 Vaccination status. A study of immunizations among school-age children in California found that some areas had rates of unvaccinated school-age children as high as 13.5% in 2014.[4] Imagine a classroom of 20 children in one such area where 13.5% of children are unvaccinated. If there are no siblings in the classroom, we are willing to consider the vaccination status of the 20 unrelated children to be independent.

a. Obtain the probability that none of the 20 children in such a classroom would be unvaccinated.
b. Obtain the probability that at least 1 of the 20 children in such a classroom would be unvaccinated.

CONDITIONAL PROBABILITY

When two events are independent, the outcome of the first event cannot influence the outcome of the second event. This is true of two independent coin tosses or two independent measurements of your height. Of course, not all events are independent. Sometimes the probability we assign to an event can change (depend on) if we know that some other event is true or has occurred. This idea is the key to many applications of probability.

EXAMPLE 10.3 Characteristics of a forest

Forests are complex, evolving ecosystems. For instance, pioneer tree species can be displaced by successional species better adapted to the changing environment. Ecologists mapped a large Canadian forest plot dominated by pioneer Douglas fir with an understory of the invading successional species western hemlock and western red cedar. The following two-way table records the distribution of all 2050 trees in the plot by species and by life stage. The distinction between live and sapling trees is made for live trees taller or shorter than 1.3 meters, respectively.[5]

	Dead	**Live**	**Sapling**	**Total**
Western red cedar (RC)	0.02	0.10	0.08	0.20
Douglas fir (DF)	0.16	0.16	0.00	0.32
Western hemlock (WH)	0.23	0.21	0.04	0.48
Total	0.41	0.47	0.12	

The “Total” row and column are obtained from the probabilities in the body of the table by the addition rule. For example, the probability that a randomly selected tree is a western hemlock (WH) is

$$P(\text{WH})=P(\text{WH and dead})+P(\text{WH and live})+P(\text{WH and sapling})=0.23+0.21+0.04=0.48$$

$$\begin{aligned} P(\text{WH}) &= P(\text{WH and dead})+P(\text{WH and live})+P(\text{WH and sapling}) \\ &= 0.23+0.21+0.04=0.48 \end{aligned}$$

Now we are told that the tree selected is a sapling. That is, it is one of the 12% of all trees that fall in the “Sapling” column of the table. The probability that a tree is a western hemlock, *given the information that the tree is a sapling,* is the proportion of western hemlocks in the “Sapling” column:

$$P(\text{WH} \mid \text{sapling}) = \frac{0.04}{0.12} = 0.33$$

This is a conditional probability. You can read the bar $\mid$ as “given the information that.”

Taylor S. Kennedy/Getty Images

Although 48% of the trees in this forest are western hemlocks, only 33% of sapling trees are western hemlock. It is common sense that knowing that one event (here, the tree is a sapling) occurs may influence the probability of another event (here, the tree is a western hemlock). The example also shows how we should define *conditional probability.* The idea of a conditional probability $P(B \mid A)$ of one event B, given that another event A occurs, is the proportion *of all occurrences of* A for which B also occurs.

CONDITIONAL PROBABILITY

When $P(A) > 0$, the **conditional probability** of B, given A, is

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

The conditional probability $P(B \mid A)$ makes no sense if the event A can never occur, so we require that $P(A) > 0$ whenever we talk about $P(B \mid A)$. *Be sure to keep in mind the distinct roles of the events A and B in* $P(B \mid A)$: Event A represents the information we are given, and B is the event whose probability we are calculating. The next example emphasizes this distinction.

EXAMPLE 10.4 Characteristics of a forest

What is the conditional probability that a randomly chosen tree is a sapling, *given the information that it is a western hemlock (WH)?* Using the definition of conditional probability,

$$\begin{aligned} P(\text{sapling} \mid \text{WH}) &= \frac{P(\text{sapling and WH})}{P(\text{WH})} \\ &= \frac{0.04}{0.48} = 0.08 \end{aligned}$$

Only 8% of western hemlock trees are saplings.

Be careful not to confuse the two different conditional probabilities:

$$\begin{aligned} P(\text{WH} \mid \text{sapling}) &= 0.33 \\ P(\text{sapling} \mid \text{WH}) &= 0.08 \end{aligned}$$

The first answers the question “What proportion of sapling trees are western hemlocks?” The second answers “What proportion of western hemlock trees are saplings?’

You should also *be careful not to confuse a conditional probability with a joint probability*. The joint probability $P(\text{sapling and WH}) = 0.04$ answers the question “What proportion of all trees in this forest are both saplings and western hemlocks?”

If you have covered optional Chapter 5 on two-way tables, you will immediately notice how similar conditional probabilities are to the conditional distributions in two-way tables.

Personal probability, defined in Chapter 9, is a special case of conditional probability. Different persons may assign different personal probabilities to a given event because their knowledge and beliefs are different. We also often update the probability we give to a particular event when new information is provided to us (or at least we should—failure to do so is known as “stubbornness”!).

APPLY YOUR KNOWLEDGE

10.14 Composition of a forest. Go back to Example 10.3. What do the data say about the current composition and the evolution of this Canadian forest?

a. Compute the conditional probabilities that a randomly chosen tree is dead, given information about its species, for each of the three tree species in this forest. What do these probabilities say about the past of this forest?

b. Compute the conditional probabilities that a randomly chosen tree is a sapling, given information about its species, for each of the three tree species. What does this say about the likely future of this forest? Which tree species appears to be taking over and becoming dominant?

10.5 Diabetes and chronic kidney disease. Diabetes and chronic kidney disease (CKD) are two diseases that are increasingly burdening the senior population in the United States. A large national sample of health care visits to the Veteran’s Health Administration system indicates that 9.1% of veterans in their 60s have stage 3–5 CKD. However, this rate is 14.1% among veterans in their 60s diagnosed with diabetes, and only 6.4% among veterans in their 60s without diabetes.[6] Express these three percents as probabilities for a randomly selected veteran 60 to 69 years of age. Use proper probability notation as in Example 10.3.

10.6 Mental health. The National Survey on Drug Use and Health reports that 18.1% of all adults in the United States had a mental illness in 2014. Among adults with a substance use disorder, 39.1% had a mental illness. By comparison, only 16.2% of adults without a substance use disorder had a mental illness. The report also states that 3.3% of American adults had both a mental illness and a substance use disorder.[7] Use the notation MI and SUD for mental illness and substance abuse disorder, respectively.

a. Express the four percents cited here as probabilities for a randomly selected American adult. Use proper probability notation as in Example 10.3.

b. Obtain the probability $P(\text{SUD}|\text{MI})$. Write a sentence reporting this probability in context.

c. Explain in context why $P(\text{SUD}|\text{MI})$ and $P(\text{MI}|\text{SUD})$ are not the same thing.

GENERAL PROBABILITY RULES

General addition rule We know that if A and B$A \text{ and } B$ are disjoint events, then P(A or B)=P(A)+P(B).$P(A \text{ or } B) = P(A) + P(B)$. If events A and B$A \text{ and } B$ are *not* disjoint, they can occur together. The probability that one or the other occurs is then *less* than the sum of their probabilities. As Figure 10.3 illustrates, outcomes common to both are counted twice when we add probabilities, so we must subtract this probability once. Here is the addition rule for any two events, disjoint or not.

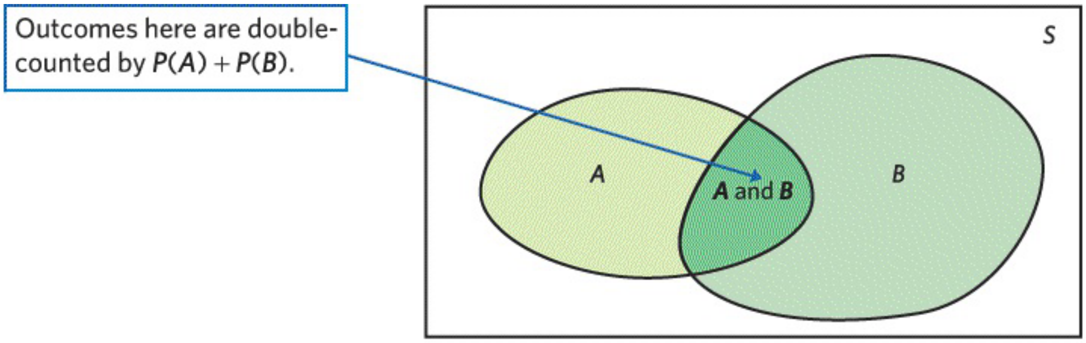


Figure 10.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 10.3 The general addition rule: P(A or B)=P(A)+P(B)-P(A and B)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

for any two events A and B.

$$A \text{ and } B.$$

GENERAL ADDITION RULE FOR ANY TWO EVENTS

For any two events A and B,$A \text{ and } B$,

P(A or B)=P(A)+P(B)-P(A and B)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If A and B are disjoint, the event $\{A \text{ and } B\}$ that both occur contains no outcomes and, therefore, has probability 0. Thus, the general addition rule includes Rule 3, the addition rule for disjoint events.

EXAMPLE 10.5 Hearing impairment in Dalmatians

Congenital sensorineural deafness is the most common form of deafness in dogs and is often associated with congenital pigmentation deficiencies. A study of hearing impairment in dogs examined thousands of Dalmatians for both hearing impairment and iris color. “Impaired” was defined as deafness in either one or both ears. Dogs with one or both blue irises (a trait due to low iris pigmentation) were labeled “blue.”

The study found that 28% of the Dalmatians were hearing-impaired, 11% were blue-eyed, and 5% were hearing-impaired and blue-eyed.[8] Choose a Dalmatian at random. Then

$$P(\text{blue or impaired})=P(\text{blue})+P(\text{impaired})-P(\text{blue impaired})=0.11+0.28-0.05=0.34$$

$$\begin{aligned} P(\text{blue or impaired}) &= P(\text{blue}) + P(\text{impaired}) - P(\text{blue impaired}) \\ &= 0.11 + 0.28 - 0.05 = 0.34 \end{aligned}$$

That is, 34% of Dalmatians were either blue-eyed or hearing-impaired. A Dalmatian is an unimpaired, brown-eyed dog if it is *neither* hearing-impaired nor blue-eyed. So

$$P(\text{neither})=P(\text{unimpaired brown})=1-0.34=0.66$$

$$P(\text{neither}) = P(\text{unimpaired brown}) = 1 - 0.34 = 0.66$$

Petra Wegner/Alamy Stock Photo

Venn diagrams are a great help in finding such probabilities because you can just think of adding and subtracting areas. Figure 10.4 shows two

Venn diagrams based on the information from Example 10.5. Both diagrams show some events and their corresponding probabilities. Venn diagram (a) displays the information we were provided: The labels B and I $B \text{ and } I$ represent all blue-eyed Dalmatians and all hearing-impaired Dalmatians, respectively. Because B and I$B \text{ and } I$ are not disjoint events, the probabilities displayed in (a) do not add to 1. In contrast, Venn diagram (b) labels the four disjoint events that make up the sample space, so the corresponding four probabilities do add to 1 (but you need to compute most of these values before you can complete the diagram). Always check the labels in a Venn diagram carefully so that you do not confuse what the probabilities displayed represent.

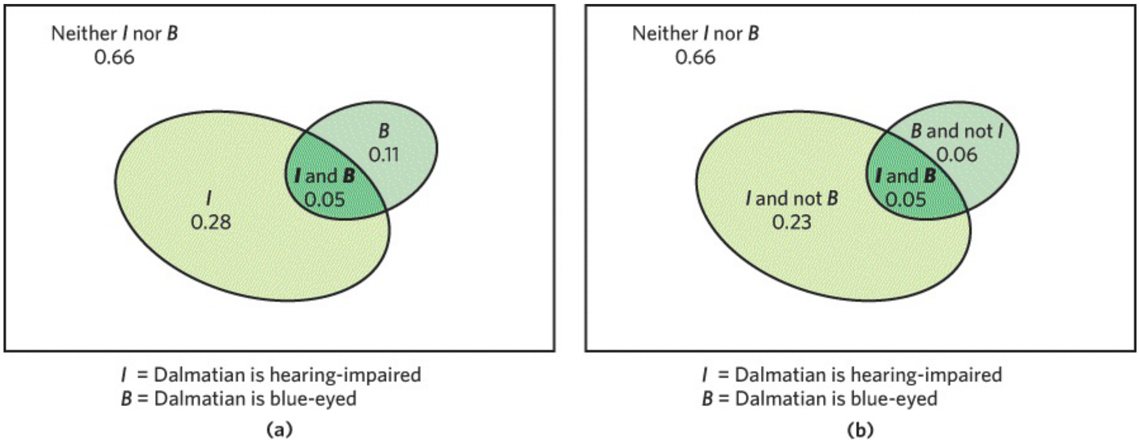

Neither I nor B
0.66
B
0.11
I and B
0.05
I
0.28
I = Dalmatian is hearing-impaired
B = Dalmatian is blue-eyed
(a)
Neither I nor B
0.66
B and not I
0.06
I and B
0.05
I and not B
0.23
I = Dalmatian is hearing-impaired
B = Dalmatian is blue-eyed
(b)

Figure 10.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 10.4 Venn diagrams and probabilities for hearing impairment and eye color in Dalmatian dogs. (a) This Venn diagram displays the simple probabilities of I and B.

$I \text{ and } B.$

(b) This Venn diagram displays the probabilities for all possible combinations of I and B.

I and B.

Notice that, in both diagrams, the probability values shown match their corresponding labels.

What is the probability that a randomly chosen Dalmatian is a hearing-impaired, brown-eyed dog? The Venn diagram in Figure 10.4(b) shows the answer directly (I and not B,0.23).$(I \text{ and not } B, 0.23)$. Figure 10.4(a) shows that this is the probability that the dog is hearing-impaired minus the probability that it is a hearing-impaired, blue-eyed dog, 0.28-0.05=0.23.

$0.28 - 0.05 = 0.23.$

Alternatively, you could display the probabilities given in Example 10.5 using a two-way table like the one in Example 10.3. Start with the information you were provided (the values shown in bold), and complete the rest of the table using the addition rule. Here is what we get:

	B	**Not *B***	**Total**
I	**0.05**	0.23	**0.28**
Not *I*	0.06	0.66	0.72
Total	**0.11**	0.89	

Some students have a strong preference for either Venn diagrams or two-way tables. You should be able to understand and use both formats. When given the choice, use the one that you find more natural and helpful.

General multiplication rule The definition of conditional probability reminds us that in principle all probabilities, including conditional probabilities, can be found from the assignment of probabilities to events that make up a random phenomenon. More often, however, conditional probabilities are part of the information given to us in a probability model. The definition of conditional probability then turns into a rule for finding the probability that both of two events occur.

GENERAL MULTIPLICATION RULE FOR ANY TWO EVENTS

The probability that both of two events A and B happen together can be found by

$$P(A \text{ and } B) = P(A)P(B \mid A)$$

$$P(A \text{ and } B) = P(A)\,P(B \mid A)$$

Here $P(B \mid A)$ is the conditional probability that B occurs, given the information that A occurs.

In words, you could think of this rule as saying that for both of two events to occur, one event must occur and, given that this event has occurred, the other event must also occur. This is often just common sense expressed in the language of probability, as the following example illustrates.

EXAMPLE 10.6 Soda consumption and weight

In a 2012 Gallup survey of a random sample of American adults, 52% said they do not drink any soda on a typical day. Of those who said they do not drink soda on a typical day, 39% described themselves as overweight. Let's consider this sample survey to be representative of the U.S. adult population. What percent of American adults drink no soda *and* are overweight?

Use the general multiplication rule:

$$P(\text{no soda})=0.52 \quad P(\text{overweight} \mid \text{no soda})=0.39 \quad P(\text{no soda and overweight})=P(\text{no soda})\times P(\text{overweight} \mid \text{no soda})=(0.52)(0.39)=0.2028$$

$$\begin{aligned} P(\text{no soda}) &= 0.52 \\ P(\text{overweight} \mid \text{no soda}) &= 0.39 \\ P(\text{no soda and overweight}) &= P(\text{no soda}) \times P(\text{overweight} \mid \text{no soda}) \\ &= (0.52)(0.39) = 0.2028 \end{aligned}$$

That is, approximately 20% of all American adults are overweight and drink no soda on typical days.

You should think your way through this equation: If 52% of Americans drink no soda and 39% *of these* people are overweight, then

> 39% of 52% are overweight and drink no soda.

The conditional probability $P(B \mid A)$ is generally not equal to the unconditional probability $P(B)$, because the occurrence of event A generally gives us some additional information about whether or not event B occurs. If knowing that A occurs gives no additional information about B, then A and B are independent events. The precise definition of independence is expressed in terms of conditional probability.

INDEPENDENT EVENTS

Two events A and B that both have positive probability are **independent** if

$$P(B \mid A)=P(B)$$

$P(B \mid A)=P(B)$

Because of the multiplication rule, this also implies that

$$P(A \text{ and } B)=P(A)P(B)$$

$P(A \text{ and } B)=P(A)P(B)$

You can use either formula to verify independence.

Looking at these formulas, we now see that the multiplication rule for independent events, $P(A \text{ and } B)=P(A)P(B)$, is a special case of the general multiplication rule, $P(A \text{ and } B)=P(A)P(B \mid A)$, just as the addition rule for disjoint events is a special case of the general addition rule.

The multiplication rule extends to the probability that all of severalevents occur. The key is to condition each event on the occurrence of *all* the preceding events. For example, for three events $A, B, \text{ and } C$:

$$P(A \text{ and } B \text{ and } C)=P(A)P(B \mid A)P(C \mid A \text{ and } B)$$

$P(A \text{ and } B \text{ and } C)=P(A)P(B \mid A)P(C \mid A \text{ and } B)$

EXAMPLE 10.7 Smoking among U.S. adults

Based on the National Health Interview Survey, the Centers for Disease Control and Prevention (CDC) estimates that 16.8% of all American adults were current smokers in 2014; that is, they smoked cigarettes either every day or on some days at the time of the interview. Daily smokers represented 76.8% of all current smokers, and 34.3% of all daily current smokers were heavy smokers (at least one pack of 20 cigarettes per day).[9]

Let's first express this information about smoking using probability notation, for the sample space representing the population of American adults in 2014:

$$P(\text{current})=0.168 P(\text{daily} \mid \text{current})=0.768 P(\text{heavy} \mid \text{current and daily})=0.343$$

$$\begin{aligned} P(\text{current}) &= 0.168 \\ P(\text{daily} \mid \text{current}) &= 0.768 \\ P(\text{heavy} \mid \text{current and daily}) &= 0.343 \end{aligned}$$

What percent of all American adults were current, daily smokers in 2014? We do not have this information, but we know what percent were current smokers *and* what percent of current smokers smoked daily. So we use the general multiplication rule:

$$P(\text{current and daily})=P(\text{current})\times P(\text{daily} \mid \text{current})=(0.168)(0.768)=0.1290$$

$$\begin{aligned} P(\text{current and daily}) &= P(\text{current}) \times P(\text{daily} \mid \text{current}) \\ &= (0.168)(0.768) = 0.1290 \end{aligned}$$

That is, approximately 12.9% of all American adults in 2014 were current smokers and smoked daily.

To find the probability that a randomly selected American adult was a current, daily, and heavy smoker in 2014, we use the multiplication rule with three events:

$$P(\text{current and daily and heavy})=P(\text{current})P(\text{daily} \mid \text{current})P(\text{heavy} \mid \text{current and daily})=(0.168)(0.768)(0.343)=0.0443$$

$$\begin{aligned}&P(\text{current and daily and heavy})\\&\quad = P(\text{current})\,P(\text{daily} \mid \text{current})\,P(\text{heavy}\mid\text{current and daily})\\&\quad = (0.168)\,(0.768)\,(0.343) = 0.0443\end{aligned}$$

That is, approximately 4.4% percent of the entire American adult population smoked at least one pack of cigarettes per day every day in 2014.

Oscar Tarragó, M.D., M.P.H., C.H.E.S./CDC

The special multiplication rule P(A and B)=P(A)P(B) $P(A \text{ and } B) = P(A)\,P(B)$ *applies only to independent events; you cannot use it when events are not independent.* Example 10.8 is a distressing example of misuse of the multiplication rule.

CAUTION

EXAMPLE 10.8 Sudden infant death syndrome

Sudden infant death syndrome (SIDS) causes babies to die suddenly (often in their cribs) with no explanation. Deaths from SIDS have been greatly reduced by placing babies on their backs to sleep, but as yet no cause for SIDS has been determined.

When more than one SIDS death occurs in a family, the parents are sometimes accused of murder. One “expert witness” (a pediatrician) popular with prosecutors in England told juries that there is only a 1 in 73 million chance that two children in the same family could have died naturally. Here’s his calculation: The rate of SIDS in a nonsmoking middle-class family is 1 in 8500. So the probability of two deaths is

18500×18500=172,250,000

$$\frac{1}{8500}\times\frac{1}{8500}=\frac{1}{72,250,000}$$

Several women were convicted of murder on this basis, without any direct evidence that they harmed their children.

The Royal Statistical Society described this reasoning as nonsense. The argument assumes that SIDS deaths in the same family are independent events. However, the cause of SIDS is unknown: “There may well be unknown genetic or environmental factors that predispose families to SIDS, so that a second case within the family becomes much more likely.”[10] As a result, the British government decided, years later, to review the cases of 258 parents convicted of murdering their babies when no direct proof was available.

Independence status (yes or no) is often part of the information given to us in a probability model or a reasonable assumption, but it can also be an important property that we want to identify. For example, lack of independence in biology can suggest a shared mechanism of action or transmission, genetic linkage, or interaction between two factors.

EXAMPLE 10.9 Dalmatians again

Example 10.5 described the frequencies of hearing impairment and eye color in Dalmatians. We know that there is a 28% chance that a randomly chosen Dalmatian is hearing-impaired, an 11% chance that it is blue-eyed, and a 5% chance that it is hearing-impaired *and* blue-eyed. We want to know if the two traits, hearing impairment and eye color, are independent or if they are linked (dependent) in Dalmatians.

We know that

$$P(\text{blue and impaired})=0.05$$

$$P(\text{blue and impaired})=0.05$$

If both events are independent, we should find the same value by using the multiplication rule for independent events:

$$P(\text{blue})\times P(\text{impaired})=0.11\times 0.28=0.03$$

$$P(\text{blue})\times P(\text{impaired})=0.11\times 0.28=0.03$$

> The true frequency of hearing-impaired, blue-eyed Dalmatians (0.05) is larger than we would predict if the two traits were independent (0.03). Therefore, hearing impairment and blue eye pigmentation in Dalmatians are not independent; they are dependent. Eye color is a well-known genetic trait. The lack of independence suggests two possible biological interpretations: Hearing impairment may also be genetically inherited in Dalmatians; or the lack of pigmentation may, in some indirect way, render the dogs more susceptible to hearing loss (for example, following a microbial infection).

APPLY YOUR KNOWLEDGE

10.7 Treating diabetes. Insulin injections (I)(I) and oral medication (OM) (OM) are two key options in diabetes treatment that may be used alone or in combination. According to the 2014 National Diabetes Statistics Report, American adults diagnosed with diabetes receive these treatments with the following probabilities:[11]

	OM	**no *OM***
I	0.147	0.140
no *I*	0.569	0.144

a. Make a Venn diagram with the information provided. Consider your options when labeling the diagram, as illustrated in Figure 10.4.

b. Obtain the probability P(I)$P(I)$ that a randomly selected American adult diagnosed with diabetes is treated with insulin, either alone or combined with oral medication. Obtain the probability P(OM) $P(OM)$ that a randomly selected American adult diagnosed with diabetes is treated with oral medication, either alone or combined with insulin.

c. What is the probability that a randomly selected American adult diagnosed with diabetes is treated with either insulin or oral medication?

d. Compute the product $P(I)\times P(OM)$$P(I) \times P(OM)$ and compare this value to the probability $P(I \text{ and } OM)$.$P(I \text{ and } OM)$. Are the insulin and oral medication treatments administered independently? Explain your answer.

10.8 Sickle-cell trait and malaria. Sickle-cell anemia is a hereditary medical condition that affects red blood cells and is thought to protect against malaria, a debilitating parasitic infection of the liver and blood. This protection would explain why the sickle-cell trait is found in people who originally came from Africa, where malaria is widespread. A landmark study in Africa tested children for the sickle-cell trait and also for malaria infection. In all, 25% of the children had the sickle-cell trait, and 6.6% of the children had both the sickle-cell trait and malaria. Overall, 34.6% of the children had malaria.[12]

a. Make a Venn diagram with the information provided. Consider your options when labeling the diagram, as illustrated in Figure 10.4.

b. What is the probability that a given child has either malaria or the sickle-cell trait?

c. What is the probability that a child has malaria, given that the child has the sickle-cell trait?

d. Are the events sickle-cell trait and malaria independent? What might that tell you about the relationship between the sickle-cell trait and malaria?

10.9 Aging physicians. The American Medical Association (AMA) is considering new guidelines for assessing the skills and abilities of aging physicians, rather than imposing a mandatory retirement age, in an effort to safeguard patients while preserving physicians' rights. The endeavor is prompted by the demographics of the aging physician workforce. The AMA estimates that "one of every four licensed U.S. doctors is older than 65, with two-fifths of them actively practicing."[13]

a. Write the information cited using probability notation.

b. What is the probability that a randomly selected licensed U.S. doctor is older than 65 and actively practicing?

10.10 Adolescents and depression. According to the 2014 National Survey on Drug Use and Health, 11.4% of adolescents in the United States had at least one major depressive episode (MDE) in the year prior to being surveyed. The percentage of MDE among adolescents was approximately 3 times higher for female adolescents (17.3%) than for male adolescents (5.7%).[14]

a. Consider the American adolescent population as the sample space of interest. Write the three percents cited here using probability notation.

b. Is MDE independent of sex in the American adolescent population? Explain your answer using the probabilities from this exercise.

TREE DIAGRAMS

People often find it difficult to understand formal probability calculations, especially when they involve conditional probabilities. The discussion (page 262) on interpreting medical test results illustrates this problem. Graphical representations of probabilities can greatly help guide calculations.

Two-way tables display joint probabilities for combinations of events, which can simply be added up to obtain marginal probabilities. Venn diagrams show these probabilities as elements scaled to represent their respective probability values. Such displays are therefore best suited when a problem provides these types of probabilities.

However, when the information provided in a problem contains conditional probabilities, computations would need to be performed before a two-way table or a Venn diagram can be completed. Some problems also have too many stages to be easily represented in a two-way table or a Venn diagram. Fortunately, we have another graphical tool to display such problems: *tree diagrams*. Here is an example.

EXAMPLE 10.10 Skin cancer in men and women

STATE: Intense or repetitive sun exposure can lead to skin cancer. Patterns of sun exposure, however, differ between men and women. A study of the presence of cutaneous malignant melanoma at a single body location among the Italian population found that 15% of skin cancers are located on the head and neck area, another 41% on the trunk, and the remaining 44% on the limbs. Moreover, 44% of the individuals with a skin cancer on the head are men, as are 63% of those with a skin cancer on the trunk but only 20% of those with a skin cancer on the limbs.[15] What percent of all individuals with skin cancer are women?

PLAN: To use the tools of probability, restate the percents as probabilities. If we choose an individual with skin cancer at random,

$$P(\text{head})=0.15P(\text{trunk})=0.41P(\text{limbs})=0.44$$

$$\begin{aligned} P(\text{head}) &= 0.15 \\ P(\text{trunk}) &= 0.41 \\ P(\text{limbs}) &= 0.44 \end{aligned}$$

These three probabilities add to 1 because every individual with skin cancer is in one of these three groups. The percents in these groups that are men are *conditional* probabilities:

$$P(\text{man} \mid \text{head})=0.44 P(\text{man} \mid \text{trunk})=0.63 P(\text{man} \mid \text{limbs})=0.20$$

$$\begin{aligned} P(\text{man} \mid \text{head}) &= 0.44 \\ P(\text{man} \mid \text{trunk}) &= 0.63 \\ P(\text{man} \mid \text{limbs}) &= 0.20 \end{aligned}$$

Biophoto Associates/Science Source

We want to find the unconditional probability $P(\text{woman})$.$P(\text{woman})$.

SOLVE: The **tree diagram** in Figure 10.5 organizes this information. Each segment in the tree is one stage of the problem. Each complete branch shows a path through the two stages. The probability written on each segment is the conditional probability that an individual with skin cancer will follow that segment, given that he or she has reached the node from which it branches.

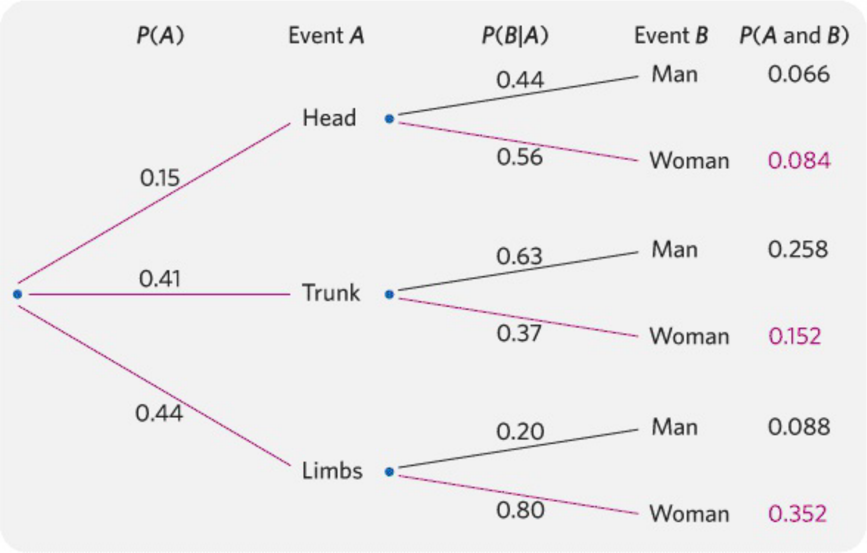


Figure 10.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 10.5 Tree diagram of skin cancer. The two stages are the location on the body of the cancer and the gender of the cancer patient.

Starting at the left, an individual with skin cancer falls into one of the three cancer location groups. The probabilities of these groups mark the leftmost branches in the tree. Look at the limbs group, the bottom branch. The two segments going out from the "limbs" branch node carry the conditional probabilities

$$P(\text{man} \mid \text{limbs})=0.20 P(\text{woman} \mid \text{limbs})=0.80$$

$$\begin{aligned} P(\text{man} \mid \text{limbs}) &= 0.20 \\ P(\text{woman} \mid \text{limbs}) &= 0.80 \end{aligned}$$

The full tree shows the probabilities for all three cancer location groups.

Now use the multiplication rule. The probability that a randomly chosen individual with skin cancer is a woman with cancer on the limbs is

$$P(\text{limbs and woman})=P(\text{limbs})P(\text{woman} \mid \text{limbs})=(0.44)(0.80)=0.352$$

$$\begin{aligned} P(\text{limbs and woman}) &= P(\text{limbs})\,P(\text{woman} \mid \text{limbs}) \\ &= (0.44)\,(0.80) = 0.352 \end{aligned}$$

This probability appears at the end of the bottom branch. You see that the probability of any complete branch in the tree is the product of the probabilities of the segments of that branch.

In this tree diagram, the event "woman" is made up of three elements: {head and woman, trunk and woman, limbs and woman}. These elements make up the three disjoint paths colored red in Figure 10.5. Because the three paths are disjoint, the probability that an individual with skin cancer is a woman is the sum of their probabilities:

P(woman)=(0.15)(0.56)+(0.41)(0.37)+(0.44)(0.80)=0.084+0.152+0.352=0.588

$$\begin{aligned} P(\text{woman}) &= (0.15)(0.56) + (0.41)(0.37) + (0.44)(0.80) \\ &= 0.084 + 0.152 + 0.352 = 0.588 \end{aligned}$$

CONCLUDE: Approximately 59% of all individuals with skin cancer are women.

It takes longer to explain a tree diagram than it does to use it. Once you have understood a problem well enough to draw the tree, the rest is easy. Here is another question about skin cancer that the tree diagram helps us answer.

EXAMPLE 10.11 Skin cancer in men and women

STATE: What percent of women with skin cancer have the cancer on their limbs?

PLAN: In probability language, we want the conditional probability

P(limbs ∣ woman)=P(limbs and woman)P(woman)

$$P(\text{limbs} \mid \text{woman}) = \frac{P(\text{limbs and woman})}{P(\text{woman})}$$

SOLVE: Look again at the tree diagram. P(woman)$P(\text{woman})$ is the overall outcome. P(limbs and woman)$P(\text{limbs and woman})$ is the result of following one branch of the tree diagram. So

P(limbs ∣ woman)=0.3520.588=0.599

$$P(\text{limbs} \mid \text{woman}) = \frac{0.352}{0.588} = 0.599$$

CONCLUDE: Approximately 60% of skin cancers in women are located on the limbs. Compare this conditional probability with the original information (unconditional) that 44% of skin cancers are located on the limbs. Knowing that a person is a woman increases the probability

> that her skin cancer is located on the limbs.

Notice how P(limbs ∣ woman) and P(woman ∣ limbs) $P(\text{limbs} \mid \text{woman})$ and $P(\text{woman} \mid \text{limbs})$ are two different probabilities. The conditional probability P(woman ∣ limbs)$P(\text{woman} \mid \text{limbs})$ is displayed on the tree and is complementary to P(man ∣ limbs)$P(\text{man} \mid \text{limbs})$—that is, they add up to 1. The conditional probability P(limbs ∣ woman),$P(\text{limbs} \mid \text{woman})$, in contrast, is calculated from other probabilities and is complementary to P(trunk ∣ woman)+P(head ∣ woman).$P(\text{trunk} \mid \text{woman}) + P(\text{head} \mid \text{woman})$.

Examples 10.10 and 10.11 illustrate a common setting for tree diagrams. Some outcome (such as the person's gender) has several sources (such as the three cancer location groups). Starting from

*⁎ the probability of each source and

*⁎ the conditional probability of the outcome, given each source,

the tree diagram leads to the overall probability of the outcome, as shown in Example 10.10. You can then use the probability of the outcome and the definition of conditional probability to find the conditional probability of one of the sources, given that the outcome occurred or is known, as shown in Example 10.11. If you like working with two-way tables better, you can use the joint probabilities at the end of the tree branches to populate the cells inside the table.

APPLY YOUR KNOWLEDGE

10.11 Eye color, hair color, and freckles. A large study of children of Caucasian descent in Germany looked at the effect of eye color, hair color, and freckles on the reported extent of burning from sun exposure.[16] The population's distribution of hair color, eye color, and freckles was as shown in the tree diagram in Figure 10.6. Find the following probabilities and describe them in plain English:

a. P(blue eyes ∣ red hair), P(blue eyes and red hair) $P(\text{blue eyes} \mid \text{red hair})$, $P(\text{blue eyes and red hair})$

b. P(freckles ∣ redhair and blue eyes), P(freckles and red hair and blue eyes)$P(\text{freckles} \mid \text{redhair and blue eyes})$, $P(\text{freckles and red hair and blue eyes})$

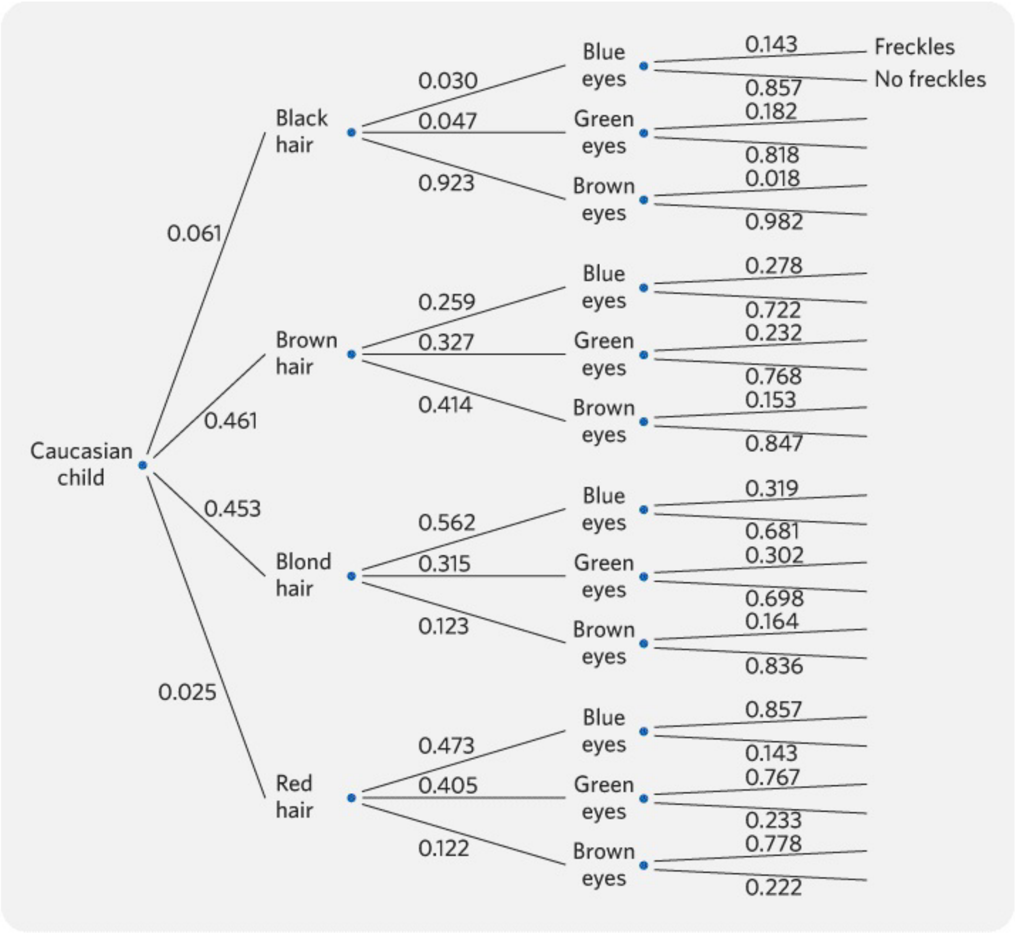

Caucasian child
0.061
0.461
0.453
0.025
Black hair
0.030
0.047
0.923
Brown hair
0.259
0.327
0.414
Blond hair
0.562
0.315
0.123
Red hair
0.473
0.405
0.122
Blue eyes
0.143
Freckles
0.857
No freckles
Green eyes
0.182
0.818
Brown eyes
0.018
0.982
Blue eyes
0.278
0.722
Green eyes
0.232
0.768
Brown eyes
0.153
0.847
Blue eyes
0.319
0.681
Green eyes
0.302
0.698
Brown eyes
0.164
0.836
Blue eyes
0.857
0.143
Green eyes
0.767
0.233
Brown eyes
0.778
0.222

Figure 10.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 10.6 Tree diagram of the features of children of Caucasian descent in Germany. The three stages are hair color, eye color, and whether or not the child has freckles.

10.12 Unintended pregnancies. Birth certificates show that approximately 9% all births in the United States are to teen mothers (ages 15 to 19), 24% to young-adult mothers (age 20 to 24), and the remaining 67% to adult mothers (age 25 to 44). An extensive survey of live births examined pregnancy type, defining an unintended pregnancy as one that was unwanted or mistimed by at least two years. The survey found that “only 23% of births to teen mothers are intended, and 77% are unintended. Among births to young-adult women age 20–24, 50% are intended, and at ages 25–44, 75% are intended.”[17]

a. Express all the percents cited here using probability notation, representing the sample space of live births in the United States.

b. Draw a tree diagram representing the age group of the mother (teen, young-adult, adult) and the pregnancy type (intended, unintended) for live births in the United States.

c. What is the probability that any given live birth in the United States is unintended?

Photographee.eu/Shutterstock.com

10.13 Eye color, hair color, and freckles, continued. Continue your work from Exercise 10.11 using the tree diagram in Figure 10.6. Find the following probabilities and describe them in plain English:

a. P(red hair), P(blue eyes), P(freckles)$P\left(\text{red hair}\right), P\left(\text{blue eyes}\right), P\left(\text{freckles}\right)$

b. P(freckles and red hair), P(freckles | red hair)

$P\left(\text{freckles and red hair}\right), P\left(\text{freckles} \mid \text{red hair}\right)$

BAYES’S THEOREM

Confusing P(B|A),$P(B|A)$, P(A|B), and P(A and B)$P(A|B)$, and $P(A \text{ and } B)$ is a common mistake. Yet, these probabilities carry very different information and can have drastically different values. In some situations, we may know the conditional probability P(B|A)$P(B|A)$ but be more interested in the inverse conditional probability P(A|B).$P(A|B)$. Here is an important example.

EXAMPLE 10.12 Diagnostic tests in medicine

The performance of a diagnostic test for a given disease can be defined by two properties. The first one is the test’s ability to appropriately give a positive result when a person tested has the disease. This is the test’s **sensitivity,** or, in probability language, PP (positive test | disease). The second property is the test’s ability to come up negative when a person tested doesn’t have the disease. This is the test’s **specificity,** PP (negative test | no disease). These two values are determined experimentally.

Information about the sensitivity of a medical test can easily be found in the internet age. However, when a patient takes a diagnostic test, what we really want to know is the probability that a positive test result truly reflects the presence of the disease. This is the conditional probability of having the disease, knowing that you received a positive test result: PP (disease | positive test). It is called the **positive predictive value** (PPV) of a test.

Many patients—and indeed even some physicians[18] —assume that if you get a positive test result, then you must have the disease. That’s not true. No test is perfectly accurate, and human error in interpreting or communicating the results can also occur. Figure 10.7 shows all possible outcomes of a diagnostic test, displayed in a two-way table and in a tree diagram. The information contained in both displays is the same, but you might feel that one is more intuitive than the other. Here is what they show:

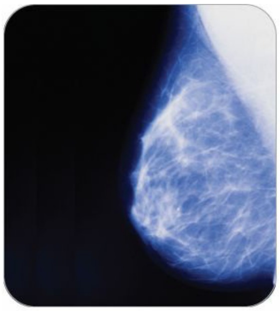
- There are four possible outcomes for a diagnostic test.
- A positive test may be obtained for subjects who have the disease (these results are called “true-positives”) and for subjects who do not have the disease (“false-positives”).

Therefore, the probability that a subject whose test result is positive truly has the disease, PP (disease | positive test), depends on what proportion of all positive tests (both true-positives and false-positives) can be expected to be true positives. This value, in turn, depends on both the properties of the diagnostic test (the test’s sensitivity and specificity) and how common or rare the disease is in the population of interest (estimated from large-scale epidemiological studies).

EXAMPLE 10.13 Mammography screening

STATE: Breast cancer occurs most frequently among older women. Of all age groups, women in their 60s have the highest rate of breast cancer. The National Cancer Institute (NCI) compiles U.S. epidemiology data for a number of different cancers. The NCI estimates that 3.65% of women in their 60s get breast cancer.[19]

Mammograms are X-ray images of the breast used to detect breast cancer. A mammogram can typically identify correctly 85% of cancer cases (sensitivity = 85%) and 95% of cases without cancer (specificity = 95%).[20] If a woman in her 60s gets a positive mammogram result, what is the probability that she really has breast cancer?



Jim Cummins/Getty Images

PLAN: We know the following probabilities:

$$P(\text{cancer})=0.0365 \text{ (cancer rate)} P(\text{no cancer})=0.9635 P(\text{test+}|\text{cancer})=0.85 \text{ (sensitivity)} P(\text{test-}|\text{cancer})=0.15 P(\text{test+}|\text{no cancer})=0.05 P(\text{test-}|\text{no cancer})=0.95 \text{ (specificity)}$$

$$\begin{aligned} P(\text{cancer}) &= 0.0365\,(\text{cancer rate}) \\ P(\text{no cancer}) &= 0.9635 \\ P(\text{test} + |\,\text{cancer}) &= 0.85\,(\text{sensitivity}) \\ P(\text{test} - |\,\text{cancer}) &= 0.15 \\ P(\text{test} + |\text{no cancer}) &= 0.05 \\ P(\text{test} - |\text{no cancer}) &= 0.95\,(\text{specificity}) \end{aligned}$$

What we want to know is

$$P(\text{cancer} \mid \text{test+})=?(\text{PPV})$$

$$P(\text{cancer} \mid \text{test} +) = ?(\text{PPV})$$

SOLVE: We can create a tree diagram like that in Figure 10.7 to visualize the problem. Figure 10.8(a) is such a tree diagram populated with the probabilities from this example. For a random woman in her 60s, we compute

$$P(\text{true-positive})=P(\text{cancer and test+})=P(\text{cancer})P(\text{test+} \mid \text{cancer})=(0.0365)(0.85)=0.031 P(\text{false-positive})=P(\text{no cancer and test+})=P(\text{no cancer})P(\text{test+} \mid \text{no cancer})=(0.9635)(0.05)=0.048$$

$$\begin{aligned} P(\text{true-positive}) &= P(\text{cancer and test+}) = P(\text{cancer})\,P(\text{test} + \mid \text{cancer}) \\ &= (0.0365)(0.85) = 0.031 \\ P(\text{false-positive}) &= P(\text{no cancer and test+}) = P(\text{no cancer})\,P(\text{test} + \mid \text{no cancer}) \\ &= (0.9635)(0.05) = 0.048 \end{aligned}$$

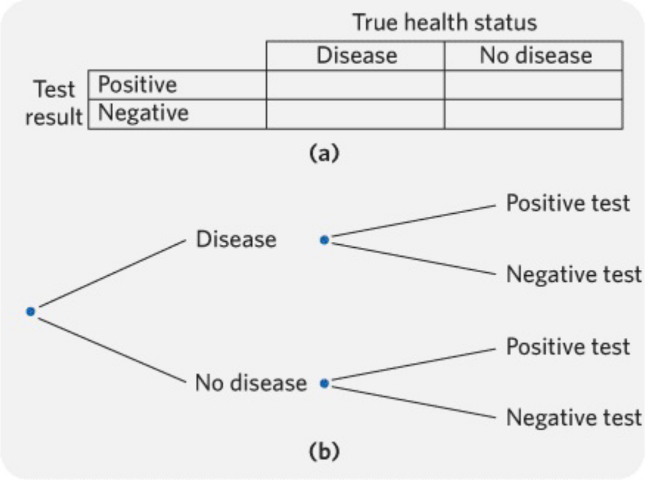
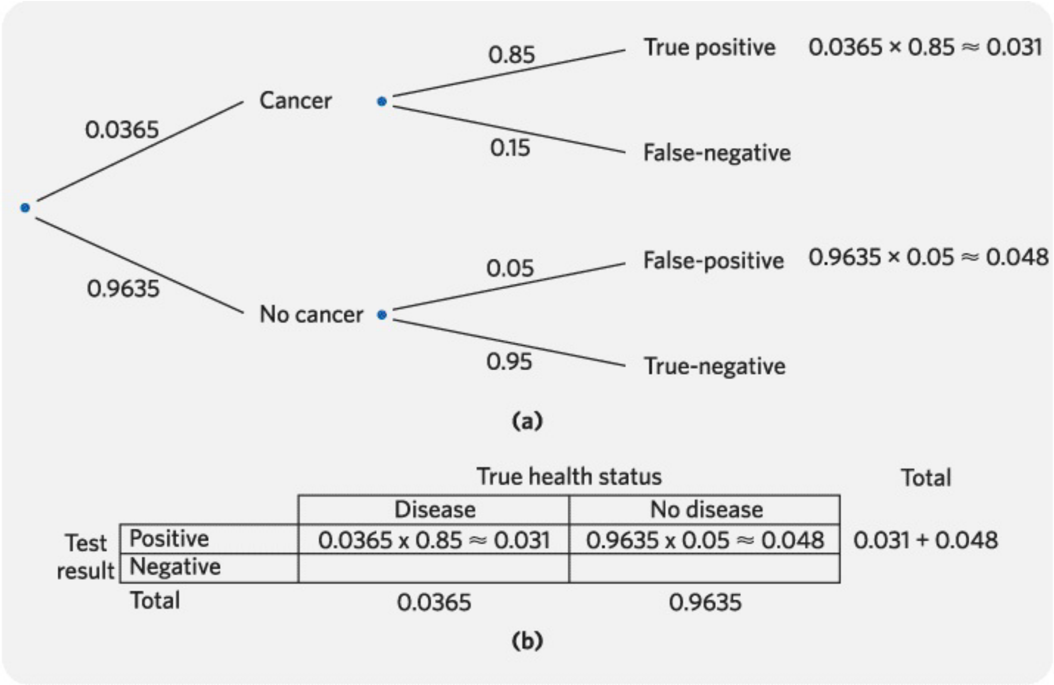


Figure 10.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 10.7 All possible outcomes of a diagnostic test, represented with (a) a two-way

table and (b) a tree diagram.


0.0365
Cancer
0.85
True positive
0.0365 × 0.85 ≈ 0.031
0.15
False-negative
0.9635
No cancer
0.05
False-positive
0.9635 × 0.05 ≈ 0.048
0.95
True-negative
(a)
(b)

		True health status		Total
		Disease	No disease	
Test result	Positive	0.0365 x 0.85 ≈ 0.031	0.9635 x 0.05 ≈ 0.048	0.031 + 0.048
	Negative			
	Total	0.0365	0.9635	

Figure 10.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 10.8 (a) Tree diagram for mammograms given to women in their 60s. (b) Corresponding two-way table.

The positive predictive value, PPV, is simply the proportion of true-positives among all positive test results for women in their 60s:

PPV=P(cancer ∣ test+)=P(true-positive)P(true-positive)+P(false-positive)=0.0310.031+0.048=0.392

$$\begin{aligned}\text{PPV} = P(\text{cancer} \mid \text{test}+) &= \frac{P(\text{true-positive})}{P(\text{true-positive})+P(\text{false-positive})} \\ &= \frac{0.031}{0.031+0.048} = 0.392\end{aligned}$$

CONCLUDE: If a randomly selected woman in her 60s gets a positive mammogram result, there is a 39% chance that she indeed has breast cancer. This also implies that there is a 61% chance that she does *not* have breast cancer. This important information should be clear to the physician and communicated clearly to the patient.

THE TRIAL OF JOHN HINCKLEY

John Hinckley attempted to assassinate President Ronald Reagan in 1981. At the trial, the defense described Hinckley's brain scan as showing brain atrophy, which could be evidence of schizophrenia. Approximately 1.5% of Americans suffer from schizophrenia. Brain atrophy is observed in 30% of schizophrenics, but only 2% of normal individuals. That is, P(atrophy ∣ schizophrenia)=0.3.$P(\text{atrophy} \mid \text{schizophrenia}) = 0.3.$ How convincing is this argument? The answer lies in the probability that Hinckley had schizophrenia, given the evidence of brain atrophy:

P(schizophrenia ∣ atrophy)=(0.3×0.015)/(0.3×0.015+0.02×0.985)=0.186.

$$\begin{aligned} P(\text{schizophrenia} \mid \text{atrophy}) &= (0.3 \times 0.015) \,/\, (0.3 \times 0.015 + 0.02 \times 0.985) \\ &= 0.186. \end{aligned}$$

What we have done in Example 10.13 is state that the probability we want (PPV) represents the frequency of one outcome (the true-positives) in a series of complementary and mutually exclusive outcomes (all positive test results: true-positives and false-positives). Using a tree diagram made it easy visually to understand why. The same information was presented in a two-way table for students who prefer working with this format.

Our approach can be summarized formally in an equation known as Bayes's theorem.

BAYES'S THEOREM

Suppose that A1,A_1, A2,A_2, …, A_k are disjoint events whose probabilities are not 0 and add to exactly 1. That is, any outcome has to be exactly in one of these events. Then if BB is any other event whose probability is not 0 or 1,

P(Ai ∣ B)=P(B ∣ Ai)P(Ai)P(B ∣ A1)P(A1)+P(B ∣ A2)P(A2)+⋯+P(Ak)P(B ∣ Ak)

$$P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+\cdots+P(A_k)P(B|A_k)}$$

The numerator in Bayes's theorem is always one of the terms in the sum that make up the denominator. The theorem has far-reaching implications in probability because it lets us express a probability as a function of prior knowledge (the A_i's). Let's apply Bayes's theorem to the breast cancer example.

EXAMPLE 10.14 Mammography screening: computations from Bayes's theorem

We can express all possible breast cancer outcomes as either cancer or no cancer (in the formula for Bayes's theorem, these outcomes would be A1 and A2).A_1 and A_2). These events are disjoint (that is, mutually exclusive) and complementary; therefore, their probabilities must add up to 1. Event BB is a positive test result. Using Bayes's theorem, we find

P(cancer ∣ test+)=P(B ∣ A1)P(A1)P(B ∣ A1)P(A1)+P(B ∣ A2)P(A2)=P(t
cancer)P(no cancer)=(0.85)(0.0365)(0.85)(0.0365)+(0.05)
(0.9635)=0.0310.031+0.048=0.392

$$
\begin{aligned}
P(\text{cancer} \mid \text{test}+) &= \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)} \\
&= \frac{P(\text{test}+|\text{cancer})P(\text{cancer})}{P(\text{test}+|\text{cancer})P(\text{cancer})+P(\text{test}+|\text{no cancer})P(\text{no cancer})} \\
&= \frac{(0.85)(0.0365)}{(0.85)(0.0365)+(0.05)(0.9635)} \\
&= \frac{0.031}{0.031+0.048} = 0.392
\end{aligned}
$$

Bayes's theorem is extremely powerful and extensively used in more advanced statistics. The importance of Bayes's theorem justifies its inclusion in an introductory textbook, but you may find its formal use challenging. It is far better to think your way through such problems rather than memorize these formal expressions without fully understanding them. The discussion that follows illustrates the benefit of this approach in the context of medical diagnostic tests.

APPLY YOUR KNOWLEDGE

10.14 Mammography screening, continued. In Examples 10.13 and 10.14 we obtained the positive predictive value, PPV=P(cancer ∣ test+),$\text{PPV} = P(\text{cancer} \mid \text{test} +)$, when mammography is used to screen asymptomatic women in their 60s. In 2009, the U.S. Preventive Services Task Force recommended against routine mammography screening for women in their 40s. In this age group, the NCI estimates that 1.49% of women get breast cancer.

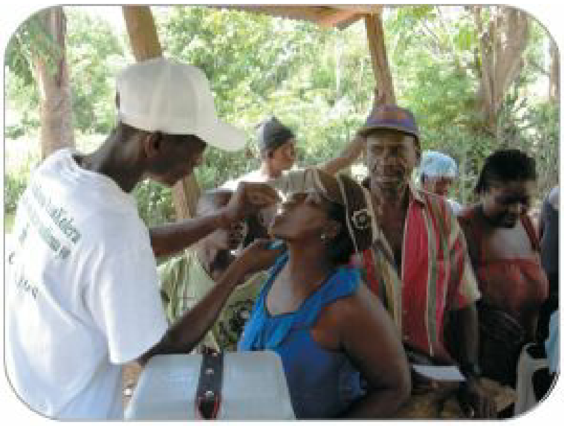
a. Examine Figure 10.8. Which parts of this figure change when the target population is women in their 40s?
b. Obtain the PPV of mammography screening for women in their 40s. How does it compare to the PPV for women in their 60s? Why do they differ?
c. Mammography screening appears to have more benefits in some age groups than in others. For women in their 40s, the evidence of a benefit is inconsistent. In addition, screening quite often results in harm from overdiagnosis, such as unnecessary biopsies and surgical treatment for nonexistent or nonthreatening cancers, for the many women who receive a positive test.[21] Explain in your own words how the information about the potential benefits and harms of mammography screening, combined with the PPV of mammography screening for women in their 40s, would lead to recommendations against routine annual screenings in this age group.

10.15 Rapid testing for cholera. In response to a recent cholera outbreak in Haiti, the National Public Health Laboratory in Haiti assessed the performance of a rapid diagnostic test for cholera. The sensitivity, PP(positive test|cholera), was found to be 95%. The specificity, PP(negative test|no cholera), was found to be 80%. It is estimated that approximately 6% of the Haitian population may have cholera.[22]

a. Draw a tree diagram representing the possible outcomes of using this cholera test in Haiti.
b. Consider testing everyone in Haiti. What percent of all tests would

be positive?

c. If a Haitian receives a positive test, what is the probability that he or she really has cholera?



Benjamin Dahl, Ph.D., M.P.H./CDC

DISCUSSION Making sense of conditional probabilities in diagnostic tests

Conditional probabilities are often grossly misinterpreted. Yet they are very important, particularly in medicine, where understanding test results requires understanding these probabilities. Many studies have documented that most physicians understand and communicate the true meaning of test results incorrectly. Here is a revealing quote from one such study:

> *For instance, doctors with an average of 14 years of professional experience were asked to imagine using the Haemoccult test to screen for colorectal cancer. The prevalence [rate] of cancer was 0.3%, the sensitivity of the test was 50%, and the false positive rate was 3%. The doctors were asked: what is the probability that someone who tests positive actually has colorectal cancer? The correct answer is about 5%. However, the doctors' answers ranged from 1% to 99%, with about half of them estimating the probability as 50% (the sensitivity) or 47% (sensitivity minus false positive rate). If patients knew about this degree of variability and statistical innumeracy they would be justly alarmed.*[23]

The positive predictive value, PPV, describes the probability that a person who receives a positive test result actually has a given disease, P(disease ∣ positive test).$P(\text{disease} \mid \text{positive test})$. You can check that PPV in this study is indeed only

(0.003)(0.5)(0.003)(0.5)+(0.997)(0.03)=0.048, or approximately 5%

$$\frac{(0.003)(0.5)}{(0.003)(0.5)+(0.997)(0.03)} = 0.048, \text{ or approximately } 5\%$$

The discrepancy between a perceived (incorrect) PPV and its actual value is particularly large for screening tests because they are typically administered to a large population with a relatively low rate of disease. When a disease is rare in the population screened, the percent of false-positives tends to be large, even when sensitivity and specificity are high. The consequence is a low PPV in the context of *indiscriminate* screening. (Note that these testing procedures have a much higher PPV when they are used to screen higher-risk individuals or to diagnose patients with a suspected condition based on existing symptoms because such groups have a higher disease prevalence.)

The potential merit of screening is that early detection may play a key role in patient prognosis. However, the emotional, medical, and financial consequences of misinterpreted test results cannot be ignored. For instance, up until the summer of 2012, the U.S. Food and Drug Administration had refused approval to over-the-counter HIV test because of concerns over public reaction to false-positives.[24] In contrast, the 2009 recommendation by the U.S. Preventive Services Task Force against routine mammography screening for women in their 40s caused an uproar,[25] despite the fact that the PPV in this age group is in the 10% to 20% range.

Beyond debating the pros and cons of various diagnostic and screening tests, we should ask how such misunderstandings can happen. Physicians are highly trained professionals, and studies show that most do know the definitions of "sensitivity" and "specificity." The challenge seems to be interpreting conditional probabilities, particularly confusing $P(A \mid B)$$P(A \mid B)$ with $P(B \mid A)$.$P(B \mid A)$. In the study cited, many physicians confused the PPV, $P(\text{disease} \mid \text{positive test})$,$P(\text{disease} \mid \text{positive test})$, with the test's sensitivity, $P(\text{positive test} \mid \text{disease})$.$P(\text{positive test} \mid \text{disease})$. Probability concepts tend to be more abstract than natural frequencies. Studies show that reframing information in a simpler, more natural format using a concrete example dramatically increases the proportion of physicians who correctly identify the PPV. The colorectal cancer case, for example, can be reframed as follows:

> Ten thousand patients take the Haemoccult test. On average, among 10,000 patients, we expect that 30 would actually have colorectal cancer and the remaining 9970 would not. Of the 30 patients with colorectal cancer, about 15 (50%) can be expected to receive a positive test result. Of the 9970 patients without colorectal cancer, about 299 (3%) can also be expected to

> receive a positive test result. What proportion of patients who receive a positive test result can be expected to actually have colorectal cancer?

The PPV seems much more intuitive now as the proportion of positive tests that are true-positives: PPV=15/(15+299)=0.048, $\text{PPV} = 15/(15 + 299) = 0.048,$ or approximately 5%. To avoid misunderstandings, researchers advocate providing information about diagnostic and screening tests in terms of natural frequencies in addition to giving the usual sensitivity and specificity conditional probabilities. Rather than being expected to perform elaborate probability calculations on their own, physicians and patients should be given clear information about the predictive value of a test. This would improve medical practice and support informed patient decisions. Figure 10.9 for Exercise 10.48 on the PSA test illustrates a different and effective way to communicate probabilities for such tests.

Until the general public understands these statistical concepts, we will continue to pay a high medical and financial price. Here is a most unusual and interesting illustration. In 2006, a Texas judge uncovered an extensive legal scam in which screening tests were used to fuel massive class-action asbestos and silicone lawsuits. Instead of representing workers who sought them out because of existing health problems, some lawyers had advertised widely to the general public and indiscriminately tested very large numbers of people, including some with existing diagnoses for unrelated lung diseases. Unsurprisingly, this practice resulted in large numbers of false-positives—and lots of money for the lawyers and screening companies. The many workers who tested positive received a meager check in the mail and were never informed that a positive test could be a false-positive. Interestingly, the scheme was uncovered by a judge who had been a nurse earlier in her career. Judge Janis Jack was the first person, despite years of such litigation, to ever ask how the plaintiffs had been screened.[26]

APPLY YOUR KNOWLEDGE

10.16 A smartphone tool to screen for syphilis. Syphilis is an easily treated sexually transmitted disease. If left untreated during pregnancy, it can cause severe health consequences for the fetus or even death. The World Health Organization recommends that all pregnant women be screened for syphilis.

Researchers describe a new smartphone accessory that can run a full lab-quality immunoassay diagnostic test, which could be used as a low-cost screening option in areas with limited health care resources. The tool can be used to screen for syphilis with a sensitivity $P(\text{positive} \mid \text{syphilis})=0.92$$P(\text{positive} \mid \text{syphilis}) = 0.92$ and a specificity $P(\text{negative} \mid \text{no syphilis})=0.79.$$P(\text{negative} \mid \text{no syphilis}) = 0.79.$[27]

a. Syphilis rates among pregnant women vary worldwide.[28] Consider a region in which 0.5% of pregnant women experience a syphilis infection. Obtain the PPV, $P(\text{syphilis} \mid \text{positive})$, $P(\text{syphilis} \mid \text{positive})$, of the smartphone tool in this population.

b. Obtain the PPV of the smartphone tool in a population where 5% of pregnant women experience a syphilis infection.

c. Compare the PPV of the smartphone tool in these two populations. Which aspect of the Discussion on page 262 do the different PPVs illustrate?

Science Picture Co/Getty Images

10.17 Rapid testing for cholera, continued. Reframe the information presented in Exercise 10.15 in terms of the number of individuals we would expect to have cholera or not, and to receive positive or negative test results, if 10,000 Haitians took this rapid diagnostic test. Then calculate the PPV from these hypothetical counts. Use the

Haemoccult test example in the Discussion on page 262 as a guideline.

CHAPTER 10 SUMMARY

- The **conditional probability** P(B ∣ A)$P(B \mid A)$ of an event B,$B,$ given an event A,$A,$ is defined by

  P(B ∣ A)=P(A and B)P(A)

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

  when P(A)>0.$P(A) > 0.$ In practice, we sometimes find conditional probabilities from directly available information rather than from the definition.

- Events A and B$A \text{ and } B$ are **independent** if knowing that one event occurs does not change the probability we would assign to the other event—that is, P(B ∣ A)=P(B).$P(B \mid A) = P(B).$ In that case, P(A and B)=P(A)P(B).$P(A \text{ and } B) = P(A)P(B).$

- Recall from Chapter 9 that events A and B$A \text{ and } B$ are **disjoint** (mututally exclusive) if they have no outcomes in common.

- Any assignment of probability obeys these rules:

  Addition rule for disjoint events: If events A,B,C,$A, B, C,$ … are all disjoint in pairs, then

  P(at least one of these events occurs)=P(A)+P(B)+P(C)+⋯

$$P(\text{at least one of these events occurs}) = P(A) + P(B) + P(C) + \cdots$$

  Multiplication rule for independent events: If events A,B,C,… $A, B, C, \ldots$ are independent, then

  P(all the events occur)=P(A)P(B)P(C)⋯

$$P(\text{all the events occur}) = P(A)P(B)P(C)\cdots$$

  General addition rule: For any two events A and B,$A \text{ and } B,$

  P(A or B)=P(A)+P(B)-P(A and B)

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

  General multiplication rule: For any two events A and B,$A \text{ and } B,$

  P(A and B)=P(A)P(B ∣ A)

$$P(A \text{ and } B) = P(A)P(B \mid A)$$

- Bayes's theorem states that, if a series of disjoint and complementary events A1A_1 through A_k have nonzero probabilities, we can express the conditional probability of event A_i, knowing that event BB occurred (when event BB has probability not equal to 0 or 1), as

  $$P(Ai \mid B)=P(B \mid Ai)P(Ai)P(B \mid A1)P(A1)+P(B \mid A2)P(A2)+\cdots +P(Ak)P(B \mid Ak)$$

  $$P(A_i \mid B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1)+P(B|A_2)P(A_2)+\cdots+P(A_k)P(B|A_k)}$$

THIS CHAPTER IN CONTEXT

Probability models provide the important connection between the observed data and the process that generated the data, as we have seen in Chapters 6 and 7 and will further explore in Chapter 13 in the context of sampling distributions. In Chapter 14 we will examine closely how probability is the foundation of statistical inference used to answer questions and draw conclusions from our data. This concept applies to all chapters of Parts III and IV, which is why it is important to have a solid understanding of probability.

Chapter 9 introduced basic ideas and facts about probability, and this chapter expands them to consider relationships between random events. The conditional distributions discussed in Chapter 5 are an example of the computation of conditional probabilities performed in this chapter. Conditional probabilities and the concept of independence play a central role when we study the relationship between two categorical variables, and we will return to them in Chapter 22 when examining the chi-square test for two-way tables. In terms of later chapters, a key idea of this chapter is independence. Many statistical procedures make an assumption of independence when considering the observations in a data set, so it is an important assumption to fully understand.

CHECK YOUR SKILLS

10.18 The weather forecast for the weekend is a 50% chance of rain for Saturday and a 50% chance of rain for Sunday. If we assume that consecutive days are independent events, what is the probability that it rains over the weekend (either Saturday or Sunday)?

a. 0.75
b. 1.0
c. Unknown, given the information provided

10.19 In fact, weather on consecutive days depends on similar atmospheric conditions, so consecutive days are not independent events. With the weather forecast of Exercise 10.18, what is the probability that it rains over the weekend (either Saturday or Sunday)?

a. 0.75
b. 1.0
c. Unknown, given the information provided

10.20 An athlete suspected of having used steroids is given two tests that operate independently of each other. Test A has probability 0.9 of being positive if steroids have been used. Test B has probability 0.8 of being positive if steroids have been used. What is the probability that *neither* test is positive if steroids have been used?

a. 0.72
b. 0.38
c. 0.02

The culinary herb cilantro is very polarizing: Some people love it and others hate it. A large survey of 12,087 American adults of European ancestry asked whether they like or dislike the taste of cilantro. Here are the numbers of men and women in the study for each answer:[29]

	Number of men	**Number of women**
Dislikes cilantro	1632	1549

Likes cilantro	5165	3741

Exercises 10.21 to 10.24 are based on this table.

10.21 What is the approximate conditional probability that such an adult likes cilantro, given that the adult is male?

a. 0.580
b. 0.737
c. 0.760

10.22 What is the approximate conditional probability that such an adult is male, given that the adult likes cilantro?

a. 0.500
b. 0.580
c. 0.760

10.23 Choose an American adult of European ancestry at random. What is the approximate probability that this person is male and likes cilantro?

a. 0.414
b. 0.427
c. 0.580

10.24 Let LL be the event that an American adult of European ancestry likes cilantro, and MM be the event that this person is male. The proportion of individuals who like cilantro among males is expressed in probability notation as

a. P(L ∣ M).$P(L \mid M)$.
b. P(M ∣ L).$P(M \mid L)$.
c. P(L and M).$P(\mathrm{L\ and\ M})$.

10.25 The CDC states on its website that "in the United States, cigarette smoking causes about 90% of lung cancers."[30] If SS is the event "smokes cigarettes" and LL is the event "has lung cancer," then the probability 0.90 is expressed in probability notation as

a. P(S and L).$P(S \text{ and } L)$.
b. P(S ∣ L).$P(S \mid L)$.

c. $P(L \mid S)$.$P(L \mid S)$.

10.26 Of people who died in the United States in a recent year, 86% were white, 12% were black, and 2% were Asian. (These values ignore a small number of deaths among other races.) Diabetes caused 2.8% of deaths among whites, 4.4% among blacks, and 3.5% among Asians. What is the approximate probability that a randomly chosen death is a white who died of diabetes?

a. 0.107

b. 0.030

c. 0.024

10.27 Using the information in Exercise 10.26, what is the approximate probability that a randomly chosen death was due to diabetes?

a. 0.107

b. 0.030

c. 0.024

CHAPTER 10 EXERCISES

10.28 Polydactyly. Polydactyly is a fairly common congenital abnormality in which a baby is born with one or more extra fingers or toes. It is reported in about 1 child in every 500.[31] A young obstetrician celebrates her first 100 deliveries. What is the probability that she has delivered no child with polydactyly? What is the probability that she has delivered at least 1 child with polydactyly?

10.29 Tasting phenylthiocarbamide. Phenylthiocarbamide (PTC) is a molecule that tastes very bitter to those with the ability to taste it. However, in about 30% of the world population, a recessive allele causes a chemosensory deficit for PTC.[32] A researcher administers a simple PTC tasting test to 10 randomly chosen subjects. What is the probability that none of these 10 subjects is PTC chemodeficient? What is the probability that at least 1 subject is chemodeficient for PTC?

10.30 Blood types. All human blood can be “ABO-typed” as O, A, B, or AB, but the distribution of the types varies a bit among groups of people. Here are the distributions of blood types for a randomly chosen person in China and in the United States:

	O	**A**	**B**	**AB**
China probability	0.35	0.27	0.26	0.12
U.S. probability	0.45	0.40	0.11	0.04

Choose an American individual and a Chinese individual at random, independently of each other. What is the probability that both have type O blood? What is the probability that both have the same blood type?

10.31 Hair color and freckles. A large study of German children of Caucasian descent found that 46% have brown hair and that 25% have freckles.[33] Nonetheless, we can’t conclude that because (0.46)(0.25)=0.115, $(0.46)(0.25) = 0.115$, about 11.5% of German children of

Caucasian descent are brown-haired with freckles. Why not?

10.32 The Rhesus factor. Human blood is typed as O, A, B, or AB, and also as Rh-positive or Rh-negative. ABO type and Rh-factor type are independent because they are governed by different genes on different chromosomes. In the U.S. population, 84% of people are Rh-positive. Use the information about ABO type in Exercise 10.30 to give the probability distribution of blood type (ABO and Rh) for a randomly chosen American.

10.33 Tendon surgery. You have torn a tendon and are facing surgery to repair it. The surgeon explains the risks to you: Infection occurs in 3% of such operations, the repair fails in 14%, and infection and failure occur together in 1%. What percent of these operations succeed and are free of infection?

10.34 Cancer-detecting dogs. Research has shown that some biochemical markers are found exclusively in the breath of patients with lung cancer. However, no lab test can currently distinguish the breath of lung cancer patients from that of other subjects. Could dogs be trained to identify these markers in specimens of human breath, just as they can be trained to detect illegal substances or to follow a person’s scent? An experiment trained dogs to distinguish breath specimens of lung cancer patients from breath specimens of control individuals by using a food-reward training method. After the training was complete, the dogs were tested on new breath specimens without any reward or clue using a double-blind, completely randomized design. Here are the results for a random sample of 1286 breath specimens:[34]

	Breath specimen from a		
Dog test result	**Control subject**	**Cancer subject**	**Total**
Negative	708	10	718
Positive	4	564	568
Total	712	574	1286

a. The sensitivity of a diagnostic test is its ability to correctly give a positive result when a person tested has the disease, or *P*(positive test | disease). Find the sensitivity of the dog cancer-detection test for lung cancer.

b. The specificity of a diagnostic test is the conditional probability that the test comes up negative, given that the subject tested doesn't have the disease. Find the specificity of the dog cancer-detection test for lung cancer.

10.35 Deer and pine seedlings. As suburban gardeners know, deer will eat almost anything green. In a study of 871 pine seedlings at an environmental center in Ohio, researchers noted how deer damage varied with how much of the seedling was covered by thorny undergrowth:[35]

	Deer damage	
Thorny cover	**Yes**	**No**
None	60	151
< 1/3	76	158
1/3 to 2/3	44	177
> 2/3	29	176

a. What is the probability that a randomly selected seedling was damaged by deer?

b. What are the conditional probabilities that a randomly selected seedling was damaged, given each level of cover?

c. Does knowing about the amount of thorny cover on a seedling change the probability of deer damage? If so, cover and damage are not independent.

10.36 World weight situation. Obesity is a growing public health concern worldwide. Adults with a high body mass index (BMI) of 25 or greater are considered overweight or obese. The following table shows the number of adults (in millions) who are overweight or obese in countries with different income levels, based on data from the World Health Organization and the United Nations.[36]

Country income level	**Millions with high BMI**	**Millions with low BMI**	**Total**
High	549	414	963

Upper middle	612	1042	1654
Lower middle	288	1083	1371
Low	63	357	420
All	1512	2896	4408

a. What is the probability that a randomly selected adult has a high BMI?

b. What are the conditional probabilities that a randomly selected adult has a high BMI, given each country income level? Write up your findings in context.

c. Compute and compare (high BMI ∣ high-income country) and (high-income country ∣ high BMI). Explain why they are not the same. Express each probability in a sentence in context.

10.37 The geometric distributions. Osteoporosis is a disease defined as very low bone density resulting in a high risk of fracture, hospitalization, and immobilization. Seven percent of postmenopausal women suffer from osteoporosis.[37] A physician administers routine check-ups. Visits by postmenopausal women are independent. We are interested in how long the physician must wait to encounter the first postmenopausal woman with osteoporosis.

a. Considering only patient visits by postmenopausal women, the probability of encountering a patient with osteoporosis on the first visit is 7%. What is the probability that the first patient does not have osteoporosis and the second patient does?

b. What is the probability that the first two patients do not have osteoporosis and the third one does? This is the probability that the first patient with osteoporosis occurs on the third visit.

c. Now you see the pattern. What is the probability that the first patient with osteoporosis occurs on the fourth visit? On the fifth visit? Give the general result: What is the probability that the first patient with osteoporosis occurs on the kth visit?

   Comment: The distribution of the number of trials to the first "success" (in our example, a patient with osteoporosis) is called a **geometric distribution.** In this problem you have found geometric distribution probabilities when the probability of a success on each trial is 7%. The same idea works for any probability of success.

10.38 Continuous probability. Choose a point at random in a square with sides $0 \le x \le 1$ and $0 \le y \le 1$.$0 \le x \le 1$ and $0 \le y \le 1$. This means that the probability that the point falls in any region within the square is equal to the area of that region. Let $X$$X$ be the $x$$x$ coordinate and $Y$$Y$ be the $y$$y$ coordinate of the point chosen.

a. Find the probability that both X and $Y$$X$ and Y are greater than 0.5. (*Hint:* Draw a diagram of the square and shade the events $X>1/2$ and $Y>1/2$.$X > 1/2$ and $Y > 1/2$.)
b. Find the conditional probability $P(Y<1/2 \mid Y>X)$.$P(Y < 1/2 \mid Y > X)$. (*Hint:* Draw a diagram of the square and shade the events $Y<1/2$ and $Y>X$.$Y < 1/2$ and $Y > X$.)

10.39 Peanut allergies: the binomial distribution. A recent survey of more than 40,000 children found that 2% were allergic to peanuts.[38] Choose 5 children at random, and let the random variable $X$$X$ be the number in this sample who are allergic to peanuts. The possible values $X$$X$ can take are 0, 1, 2, 3, 4, and 5. Make a five-stage tree diagram of the outcomes (allergic or not allergic) for the 5 individuals, and use it to find the probability distribution of X.X. *(We will see in Chapter 12 that X follows a binomial distribution.)*

10.40 Hepatitis C and HIV coinfection. According to a 2003 report of the U.S. Health Resources and Services Administration, coinfection of HIV and hepatitis C (HCV) is on the rise. Only about 2% of Americans have HCV, but 25% of Americans with HIV have HCV. Approximately 10% of Americans with HCV also have HIV.[39]

a. Write the information provided in terms of probabilities. Are the two viral infections independent? Explain your reasoning.
b. What is the probability that a randomly chosen American has both HIV and HCV?

10.41 Tay-Sachs disease. Tay-Sachs (TS) disease is a fatal, recessive genetic disorder in which harmful quantities of a fatty waste substance accumulate in the brain, slowly progressing until it leads to death by age five. TS carriers are common among individuals of Eastern European Jewish descent, with a carrier frequency of about 1 in 27. In the general population, the carrier rate is about 1 in 250.[40]

a. Give the probability that both parents are carriers in each of the following conditions: Both mom and dad are of European Jewish descent; mom is of European Jewish descent but dad isn't; both mom and dad are not of European Jewish descent.
b. If both parents are carriers, the probability is 0.25 that their child has TS. Otherwise, the probability is zero. Draw the tree diagram for the children of parents who are both of European Jewish descent. Use the tree to find the probability that a child has TS, given that both parents are of European Jewish descent.
c. What would be the probability that a child has TS, given that neither parent is of European Jewish descent?

10.42 Albinism. The gene for albinism in humans is recessive. That is, carriers of this gene have probability 1/2 of passing it to a child, and the child is albino only if both parents pass the albinism gene. Parents pass their genes independently of each other. If both parents carry the albinism gene but are not albino, what is the probability that their first child is albino? If they have two children (who inherit their genes independently of each other), what is the probability that both are albino? That neither is albino?

10.43 Cystic fibrosis. Cystic fibrosis (CF) is a hereditary lung disorder that often results in death. It can be inherited only if both parents are carriers of an abnormal gene. In 1989, the CF gene that is abnormal in carriers of cystic fibrosis was identified. The probability that a randomly chosen person of European ancestry carries an abnormal CF gene is 1/25. (The probability is less in other ethnic groups.) The CF20m test detects most, but not all, harmful mutations of the CF gene. The test is positive for 90% of people who are carriers. It is (ignoring human error) never positive for people who are not carriers.[41] Jason's CF20m test is positive. What is the probability that he is a carrier?

10.44 Lactose intolerance. Lactose intolerance causes difficulty digesting dairy products that contain lactose (milk sugar). It is particularly common among people of African and Asian ancestry. In the United States (ignoring other groups and people who consider themselves to belong to more than one race), 82% of the population is white, 14% is black, and 4% is Asian. Moreover, 15% of whites, 70% of blacks, and 90% of Asians are lactose intolerant.[42]

a. What percent of the entire population is lactose intolerant?
b. What percent of people who are lactose intolerant are Asian?

10.45 Unintended pregnancies, continued. Go back to Exercise 10.12 about unintended pregnancies in the United States. Use your tree diagram and Bayes's theorem to find the probability that a birth was to a teen mother if we know that the pregnancy was unintended.

10.46 CT scans for heavy smokers. Lung cancer is the leading cause of cancer-related deaths in the United States. Researchers examined the idea of testing all Medicare-enrolled heavy smokers for lung cancer with a computed tomography (CT) scan every year. In this population, the lifetime chance of developing lung cancer is high. In any given year, approximately 3% of heavy smokers develop lung cancer. The CT scan positively identifies lung cancer 89% of the time, and it gives a negative result for 93% of individuals who do not have lung cancer.[43]

a. Use probability notation to express the three values cited.
b. Make a probability tree representing the outcomes in the sample space of Medicare-enrolled heavy smokers. Be sure to place probability values on all the branches of your tree.
c. What percent of CT scans in this target population would be positive?
d. What is the probability that a Medicare-enrolled heavy smoker who gets a positive CT scan actually has lung cancer, $(\text{cancer} \mid \text{positive})$?

10.47 Assessing the new OraQuick HIV test. In 2012, the U.S. Food and Drug Administration approved the first over-the-counter HIV testing kit, the OraQuick In-Home HIV Test. Example 10.12 told us that the sensitivity and specificity of a diagnostic test are determined experimentally. The OraQuick test was approved based on a study that examined the test results of a sample of 4410 adults from a high-risk population who did not yet know their HIV status. Each person's actual HIV status was later confirmed via several other medical examinations. Here are the findings:[44]

105 true-positives (HIV and positive test)	1 false-positive (no HIV and positive test)
8 false-negatives (HIV and negative test)	4296 true-negatives (no HIV and negative test)

a. What is the sensitivity, PP (positive test ∣ ∣ HIV), of OraQuick? This represents the fraction of all HIV cases who received a positive test result. Use the counts in the table of findings to obtain your answer.
b. What is the specificity, PP (negative test ∣ ∣ no HIV), of OraQuick? This represents the fraction of all individuals with no HIV who received a negative test result.
c. Now use the study findings to find the approximate rate of HIV in this high-risk population. What is the positive predictive value of OraQuick, PP (HIV ∣ ∣ positive test), in this high-risk population?
d. Explain in your own words why the positive predictive value of this test would be different for a population with a different rate of HIV, whereas the test's sensitivity and specificity would remain the same.

10.48 Should I have a PSA test? That's the question addressed in a Kaiser Permanente pamphlet on prostate cancer. The prostate-specific antigen (PSA) test is a blood test that is used to screen for prostate cancer. Elevated PSA levels are often associated with prostate cancer, but they can also reflect other conditions. Thus, a positive PSA test does not necessarily indicate prostate cancer. Moreover, not all cases of prostate cancer can be detected with a PSA test. The pamphlet uses a clever diagram to represent the possible outcomes of a PSA test with their respective frequencies.[45]

a. The Kaiser diagram is reproduced in Figure 10.9. Use this diagram to find the probabilities of a true-positive, a false-positive, a false-negative, and a true-negative.

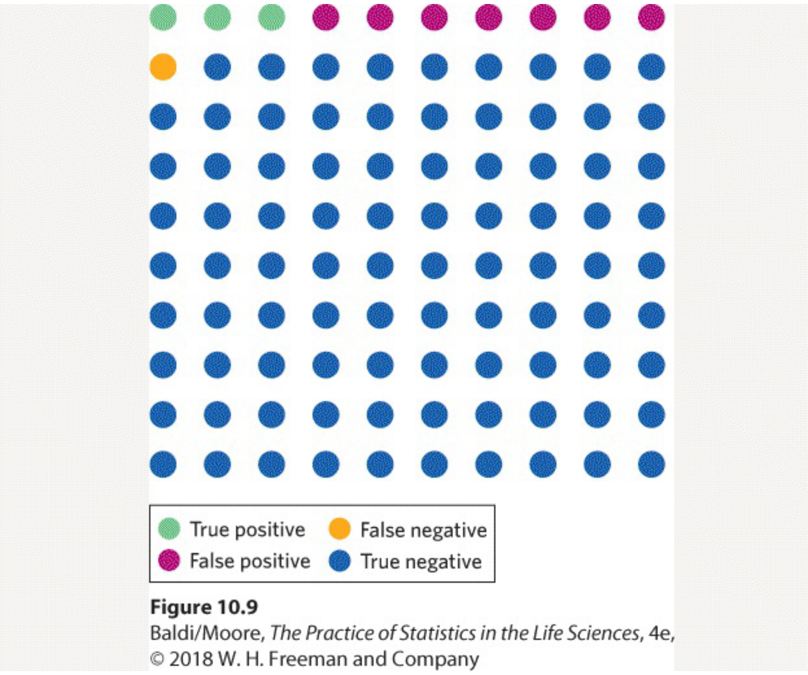


Figure 10.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 10.9 Kaiser Permanente’s outcome diagram of 100 PSA tests. Each dot represents the result of one PSA test.

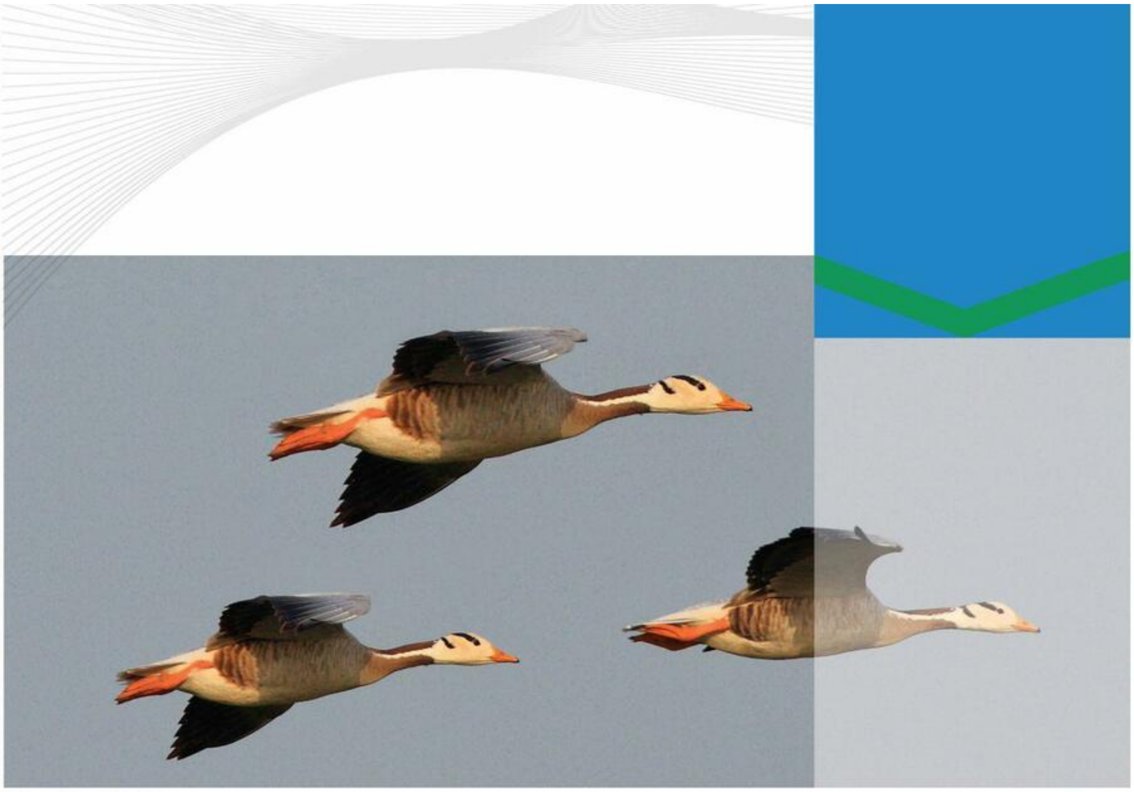
b. What is the positive predictive value (PPV) of the test? That is, what is the conditional probability that a randomly screened patient has prostate cancer, given that he receives a positive PSA test result?

c. In addressing the question “Should I have a PSA test?” Kaiser discusses the relatively low PPV of the test. The pamphlet also mentions that treatment for prostate cancer can cause serious side effects and that most men diagnosed with prostate cancer die of other causes. (Some prostate cancers are aggressive and spread, but most don’t.) The last two facts are unrelated to the reliability of the test as assessed by its PPV. Explain why they are nonetheless important considerations in deciding whether to include a PSA test as part of routine medical check-ups for men older than age 50.

10.49 Asking sensitive questions. When asked personal questions of a sensitive nature, survey respondents are not always truthful. We saw in Chapter 6 that this outcome is an issue of response bias. One way to make it easier for respondents to answer truthfully is to give them an opportunity to preserve the anonymity of their answer. For example, one randomized-response technique lets respondents flip a coin without showing anyone. If the result is a head, the participant answers the sensitive question truthfully (“yes” or “no”); otherwise, the participant just answers “yes” regardless of the truth. Therefore, no one but the respondent can tell if a “yes” was actually an answer to the sensitive question. However, when all the answers are compiled, researchers can use the data to figure out what proportion of individuals answered “yes” to the sensitive question. Here is how.

a. Let pp represent the proportion of individuals who answered “yes” to the sensitive question. Make a tree diagram representing all possible outcomes for a survey respondent, starting with the coin flip. Be sure to place the correct probabilities on your tree.
b. Use your tree diagram to compute the probability that a survey respondent would answer “yes.” Note that this information, P(yes), is readily available from the complete survey data.
c. Assume that the sensitive question was “Have you used illegal drugs in the past six months?” and that 59% of the participants answered “yes” in this randomized-response design. Using your work in part b, find the proportion pp of individuals who used illegal drugs in the past six months.

CHAPTER 11 The Normal Distributions



Nadmme Khawar/Getty Images

IN THIS CHAPTER WE COVER...

- Normal distributions
- The 68–95–99.7 rule
- The standard Normal distribution
- Finding Normal probabilities
- Finding percentiles
- Using the standard Normal table
- Normal quantile plots

Of all the continuous probability distributions, none is more widely used than the family of Normal distributions. Normal distributions are extremely important in statistical inference, as we will see starting with Chapter 14. They are also a good mathematical model for many (but certainly not all) biological variables, such as blood pressure, bone mineral density, daily global temperature, and the height of plants or animals.

NORMAL DISTRIBUTIONS

Chapter 9 introduced the idea of density curves to describe the probability distributions of continuous random variables. One particularly important class of density curves appeared in Figure 9.2. These symmetric, single-peaked, and bell-shaped density curves are called *Normal curves*, and they describe *Normal distributions*. Normal distributions play a large role in statistics. The name “normal” has a certain logic in a historical context, but it should not be taken to imply something usual, typical, or mainstream. We capitalize “Normal” to remind you that the term refers to a mathematical function rather than an assertion of ordinariness.

All Normal distributions have the same overall shape. The exact equation[1] describing a Normal distribution is beyond the scope of this textbook, but from this equation we know that *the exact density curve for a particular Normal distribution is described by giving its mean* μμ *and its standard deviation* σ.$\sigma.$ As we saw in Chapter 9, the mean and standard deviation of a random variable describe, respectively, the center and spread of its probability distribution. We use the Greek letters μμ and σσ to indicate that these are parameters summarizing all possible outcomes of the random variable (the population) rather than a sample set of observations.

The mean μμ of a Normal distribution may be any value, positive or negative, and its standard deviation σσ may be any positive value (standard deviations can, by definition, be only positive values). Figure 11.1 shows two Normal curves with different values of μμ and σ.$\sigma.$ Because a Normal curve is symmetric, its mean is located at the center of symmetry and is the same as the median. The value of the mean μμ anchors a Normal curve to a particular location on the horizontal axis. Changing μμ without changing σσ would simply move the Normal curve along the horizontal axis without changing its spread. The standard deviation σσ controls the spread of a Normal curve, such that a curve with a larger standard deviation is more spread out. Recall that density curves, by definition, are scaled so that the total area under the curve equals 1 (the probability of the entire sample space). This is why, in Figure 11.1, the Normal curve on the left with the larger standard deviation σ1σ_1 is not only wider but also shorter than the

Normal curve on the right, which has the smaller standard deviation $\sigma 2$.σ_2.

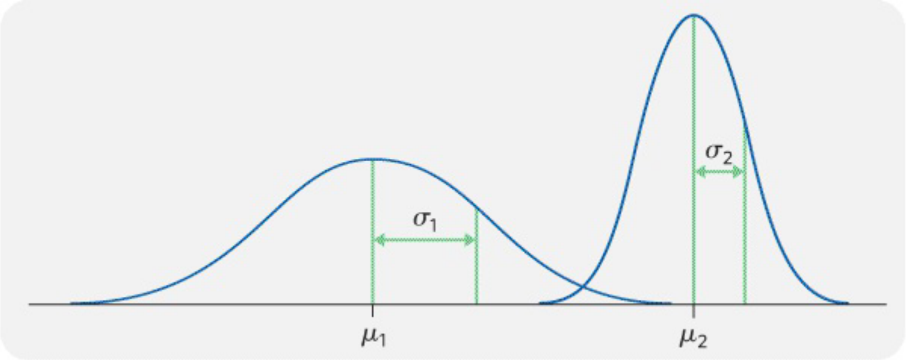
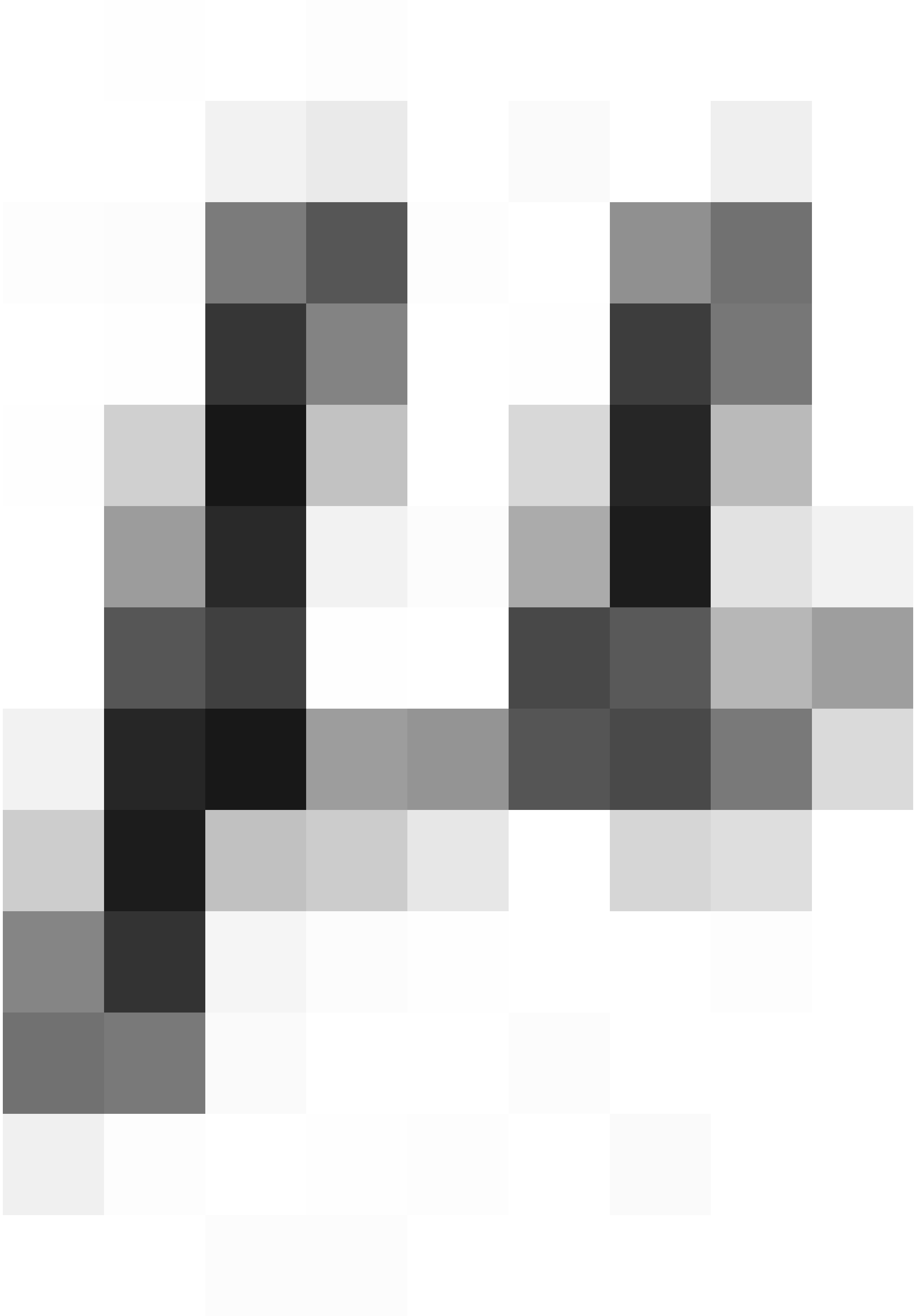


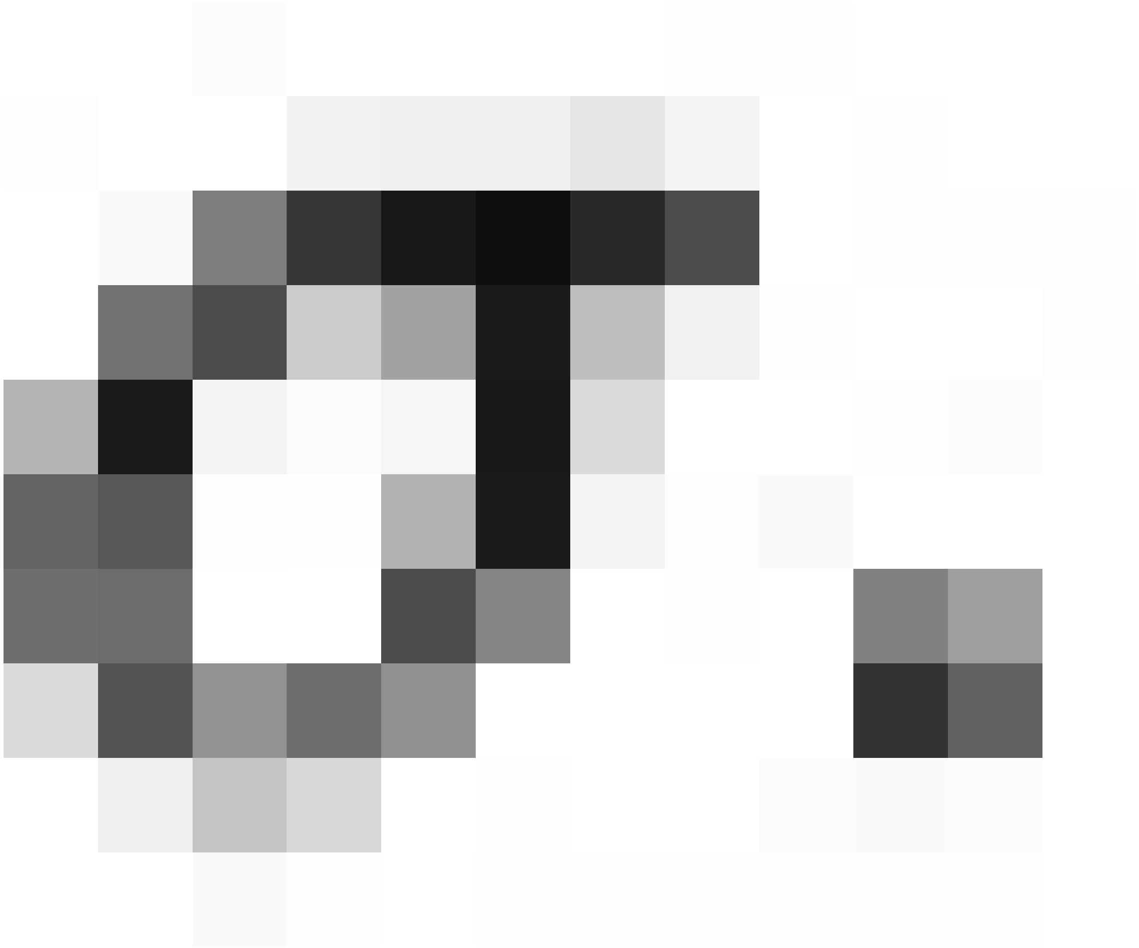
Figure 11.1

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

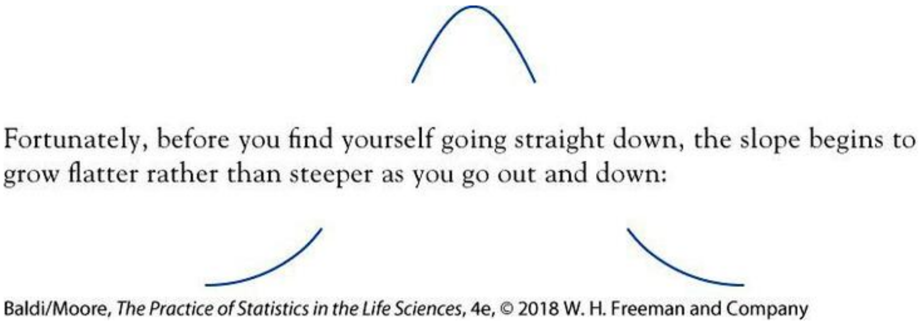
FIGURE 11.1 Two Normal curves with a different mean μ



and a different standard deviation σ.



The parameters μμ and σσ completely determine the shape of a Normal curve, and we can identify them visually on the curve. The mean μμ is located at the center of symmetry, under the tallest point of the density curve. The standard deviation σσ is related to a point of curvature change on the density curve. To get a feel for this concept, imagine that you are skiing or biking down a mountain that has the shape of a Normal curve. At first, you descend at an ever-steeper angle as you go out from the peak:



Fortunately, before you find yourself going straight down, the slope begins to grow flatter rather than steeper as you go out and down:

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

The points at which this change of curvature takes place are located at distance σ *on either side of the mean* μ. You can feel the change as you run a pencil along a Normal curve. Remember that μ *and* σ *alone do not specify the shape of most other distributions,* and that the shape of density curves in general does not reveal σ. These are special properties of Normal distributions.

NORMAL DISTRIBUTIONS

A **Normal distribution** is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers: its mean and its standard deviation.

The mean of a Normal distribution is at the center of the symmetric Normal curve. The standard deviation is the distance from the center to the change-of-curvature points on either side.

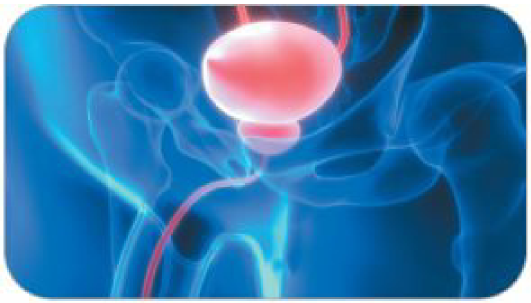
Why are the Normal distributions important in statistics? Here are three reasons. First, Normal distributions are good *models* to describe some distributions of real data. Distributions that are often close to Normal include scores on tests taken by many people (such as IQ tests and MCAT exams), repeated careful measurements of the same quantity, and characteristics of biological populations (such as lengths of crickets and yields of corn). Second, Normal distributions are good approximations for the results of some kinds of *chance outcomes*, such as the proportion of individuals diagnosed with diabetes in sample surveys of 1000 adults. Third, many *statistical inference* procedures based on Normal distributions work well for other, roughly symmetric distributions.

Students and professionals often make the error of assuming that their variable is Normally distributed without first verifying this assumption by plotting the data. However, many data sets do not follow a Normal distribution. Distributions of body weights or survival times, for example, are skewed to the right and so are not Normal. Figure 1.7 (page 18), showing the distribution of rates of pertussis for all California counties in 2014, is a clear example of non-Normal data. Keep in mind that non-Normal data are quite common, so be sure to always plot your data.

APPLY YOUR KNOWLEDGE

11.1 Bladder volumes. The distribution of bladder volumes in men is approximately Normal with mean 550 milliliters (ml) and standard deviation 100 ml. In women, bladder volumes are approximately Normal with mean 400 ml and standard deviation 75 ml.[2]

a. Draw a line and scale it from 100 to 900 ml, by increments of 100 ml. Mark the mean of each distribution on your graph. Which of the two distributions is the widest? Should the widest distribution also be the tallest or the shortest? Explain your answer.
b. Add a rough sketch of these two Normal curves on your graph, marking the change-of-curvature points. Then add the value of each distribution's standard deviation, as in Figure 11.1.



Magic mine/Shutterstock.com

11.2 Mercury levels in pregnant women. Exercise 2.4 described a study of blood mercury levels in pregnant British women. The researchers reported that "blood mercury levels ranged from 0.17 to 12.8 μg/l. The 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles were 0.81, 0.99, 1.35, 1.86, 2.52, 3.33, and 4.02 μg/l, respectively." Explain why a Normal distribution would not be an appropriate model for the distribution of blood mercury levels in pregnant British women.

THE 68–95–99.7 RULE

Although many Normal curves are possible (with different values of μμ and σσ), they all have common properties. In particular, all Normal distributions obey the following rule.

THE 68–95–99.7 RULE

In the Normal distribution with mean μμ and standard deviation σ:σ:

- Approximately **68%** of the observations fall within σσ of the mean μ.μ.
- Approximately **95%** of the observations fall within 2σ2σ of μ.μ.
- Approximately **99.7%** of the observations fall within 3σ3σ of μ. μ.

Expressed in probability terms, this rule means that if the random variable XX has a Normal distribution with mean μμ and standard deviation σ,σ, then

P(μ-σ<X<μ+σ)≈0.68P(μ-2σ<X<μ+2σ)≈0.95P(μ-3σ<X<μ+3σ)≈0.997

$$\begin{aligned} P(\mu-\sigma<X<\mu+\sigma) &\approx 0.68 \\ P(\mu-2\sigma<X<\mu+2\sigma) &\approx 0.95 \\ P(\mu-3\sigma<X<\mu+3\sigma) &\approx 0.997 \end{aligned}$$

Figure 11.2 illustrates the 68–95–99.7 rule. By remembering these three numbers, you can think about Normal distributions without constantly making detailed calculations. This rule can also be very useful to quickly check that your detailed calculations are not obviously wrong.

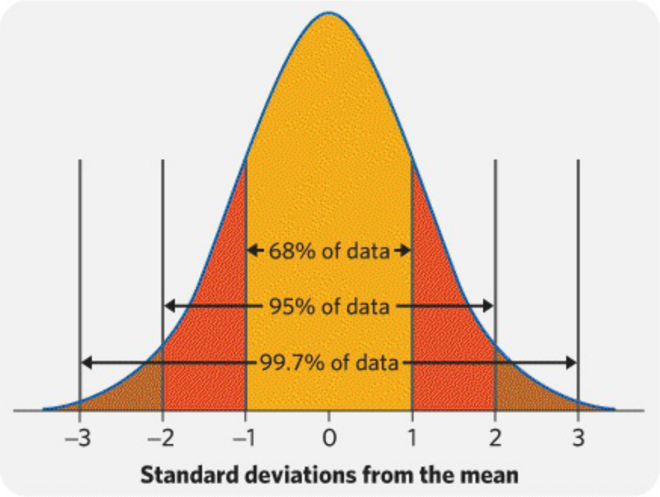


Figure 11.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 11.2 The 68–95–99.7 rule for Normal distributions.

EXAMPLE 11.1 Heights of young women

The distribution of heights of young women age 18 to 24 is approximately Normal, with mean $\mu=64.5$$\mu = 64.5$ inches and standard deviation $\sigma=2.5$ inches.$\sigma = 2.5\,\text{inches.}$ Figure 11.3 applies the 68–95–99.7 rule to this distribution.

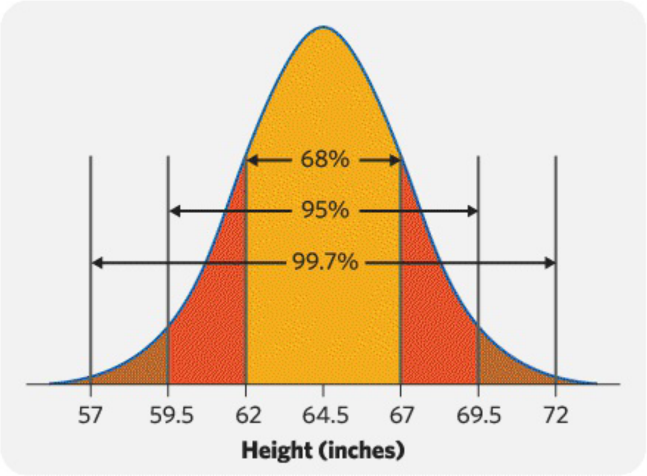


Figure 11.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 11.3 The 68–95–99.7 rule applied to the distribution of heights of young women age 18 to 24 years, with $\mu=64.5$ inches

$$\mu = 64.5 \text{ inches}$$

and $\sigma=2.5$ inches.

$$\sigma = 2.5 \text{ inches.}$$

The 95 part of the 68–95–99.7 rule says that the heights of the middle 95% of young women are approximately between

$$\mu-2\sigma=64.5-(2)(2.5)=64.5-5=59.5$$

$\mu - 2\sigma = 64.5 - (2)(2.5) = 64.5 - 5 = 59.5$

and

$$\mu+2\sigma=64.5+(2)(2.5)=64.5+5=69.5$$

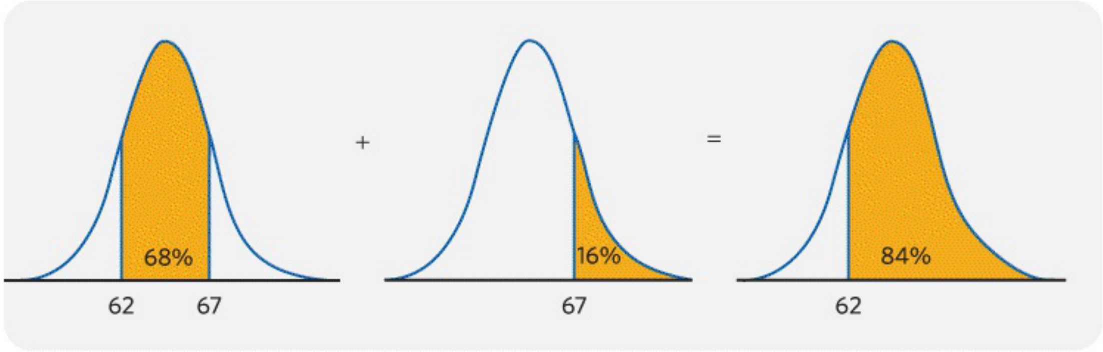
$\mu + 2\sigma = 64.5 + (2)(2.5) = 64.5 + 5 = 69.5$

That is, these women are between 59.5 and 69.5 inches tall.

The other 5% of young women have heights outside this range. Because the Normal distributions are symmetric, half of these women are on the tall side. So the tallest 2.5% of young women are taller than 69.5 inches. Rephrased in probability terms, this means that there is probability approximately 0.025 that a randomly selected woman is taller than 69.5 inches.

EXAMPLE 11.2 Heights of young women

Look again at Figure 11.3. A height of 62 inches is one standard deviation below the mean. What is the probability that a woman is taller than 62 inches? Find the answer by adding areas in the figure. Here is the calculation in pictures:



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

percent between 62 and 67+percent above 67=percent above 6268%+16%=84%

$$\begin{array}{ccccc}\text{percent between 62 and 67} & + & \text{percent above 67} & = & \text{percent above 62}\\ 68\% & + & 16\% & = & 84\%\end{array}$$

Be sure you see where the 16% came from: 32% of heights are outside the range 62 to 67 inches, and half of these are greater than 67 inches.

The 68–95–99.7 rule describes distributions that are exactly Normal. Real data are never "exactly" Normal, however. Look back at Example 9.9 (page 225) on the heights of women in their 40s. The example and Figure 9.3 show that calculations using the histogram of raw data are quite close to calculations using a Normal density curve. Differences are due, in part, to the fact that someone's height is usually reported only to the nearest quarter of an inch. In other words, a height may be reported as 59.0 or 59.25, but not 59.1436. We can use a Normal distribution as a model for heights because it provides a good approximation and because we know that heights are actually continuous—people do not grow by steps of a quarter of an inch!

Because we will mention Normal distributions often, a short notation is helpful. We abbreviate the Normal distribution with mean μμ and standard deviation σσ as N(μ,σ).$N(\mu, \sigma)$. For example, the distribution of young women's heights is approximately N(64.5, 2.5).$N(64.5,\ 2.5)$.

APPLY YOUR KNOWLEDGE

11.3 Bladder volumes, continued. Go back to Exercise 11.1 and the sketches you made for it. Use the 68–95–99.7 rule to answer the following questions.

a. Between what volumes do the middle 95% of men's bladders fall?
b. What percent of men's bladders have a volume larger than 650 ml?
c. Between what values do almost all (99.7%) of women's bladder volumes fall?
d. How small are the smallest 2.5% of all bladders among women?

11.4 Flying over the Himalayas. Bar-headed geese cross the Himalayan mountain range during their biannual migration, which means that their physiology and flight must adapt to very high altitudes (greater than 3 miles, or 4.8 kilometers) with low air density. Researchers implanted small recording instruments on wild bar-headed geese and found that wingbeat frequency at very high altitudes is roughly Normally distributed, with a mean frequency of 4.25 flaps per second (hertz, Hz) and a standard deviation of 0.2 Hz.[3]

a. Sketch the distribution of wingbeat frequency in bar-headed geese at very high altitude. Place the values of the mean and up to 3 standard deviations around the mean on the horizontal axis, as in Figure 11.3.
b. At very high altitude, between what values do almost all (99.7%) wingbeat frequencies fall? Use the 68–95–99.7 rule for your answer.
c. What is the probability that a bar-headed goose flying at very high altitude has a wingbeat frequency between 4.05 and 4.45 Hz? Use the 68–95–99.7 rule for your answer.

Nadmme Khawar/Getty Images

THE STANDARD NORMAL DISTRIBUTION

As the 68–95–99.7 rule suggests, all Normal distributions share many common properties. In fact, all Normal distributions are the same if we measure in units of size σ about the mean μ as center. Changing to these units is called *standardizing*. To standardize a value, subtract the mean of the distribution and then divide by the standard deviation.

STANDARDIZING AND *Z*-SCORES

If x is an observation from a distribution that has mean μ and standard deviation σ, the **standardized value** of x is

$$z=\frac{x-\mu}{\sigma}$$

A standardized value is often called a ***z*-score.**

A *z*-score tells us how many standard deviations the original observation falls away from the mean, and in which direction. Observations larger than the mean are positive when standardized, and observations smaller than the mean are negative.

EXAMPLE 11.3 Standardizing women's heights

The heights of young women age 18 to 24 are approximately Normal, with $\mu = 64.5 \text{ inches}$ and $\sigma = 2.5 \text{ inches}$. The standardized height is

$$z=\frac{\text{height}-64.5}{2.5}$$

A woman's standardized height is the number of standard deviations by which her height differs from the mean height of all young women. A woman 70 inches tall, for example, has standardized height

$$z=\frac{70-64.5}{2.5}=2.2$$

> $$z = \frac{70-64.5}{2.5} = 2.2$$
>
> or 2.2 standard deviations above the mean. Similarly, a woman 5 feet (60 inches) tall has standardized height
>
> z=60-64.52.5=-1.8
>
> $$z = \frac{60-64.5}{2.5} = -1.8$$
>
> or 1.8 standard deviations less than the mean height.

We often standardize observations from symmetric distributions to express them using a common scale. We might, for example, measure the height of a boy at two different ages by calculating each *z*-score. The standardized heights tell us where the child stands in the distribution for his age group and whether he might be growing more slowly than other children of his age.

If the variable we standardize has a Normal distribution, standardizing does more than give a common scale: It turns all Normal distributions into a single distribution, and this distribution is still Normal. Standardizing a variable that has any Normal distribution produces a new variable that has the *standard Normal distribution.*

STANDARD NORMAL DISTRIBUTION

The **standard Normal distribution** is the Normal distribution N(0,1)$N(0,1)$ with mean 0 and standard deviation 1.

If a variable xx has any Normal distribution N(μ,σ)$N(\mu,\sigma)$ with mean μμ and standard deviation σ,$\sigma,$ then the standardized variable

z=x-μσ

$$z = \frac{x-\mu}{\sigma}$$

has the standard Normal distribution.

Keep in mind that *standardizing a variable x by subtracting the mean and dividing by the standard deviation does not automatically make the resulting z-score Normal.* The resulting *z*-score will follow the standard Normal distribution N(0,1)$N(0,1)$ only if the original xx variable had a Normal distribution N(μ,σ).$N(\mu,\sigma)$.

APPLY YOUR KNOWLEDGE

11.5 Growth chart. Boys' growth charts indicate that the heights of five-year-old boys are approximately Normal with mean 43 inches and standard deviation 1.8 inches, whereas those of seven-year-old boys are approximately Normal with mean 48 inches and standard deviation 2.1 inches. John measured 44 inches on his fifth birthday. Now, on his seventh birthday, he measures 49 inches. Find John's standardized scores for both ages and compare them. Has he been growing faster or more slowly than the general population of boys his age?

11.6 Men's and women's heights. The heights of women age 20 to 29 are approximately Normal, with mean 64 inches and standard deviation 2.7 inches. Men the same age have mean height 69.3 inches, with standard deviation 2.8 inches. What are the *z*-scores for a woman 6 feet tall (72 inches) and for a man 6 feet tall? Say in simple language what information the *z*-scores give that the actual heights do not.

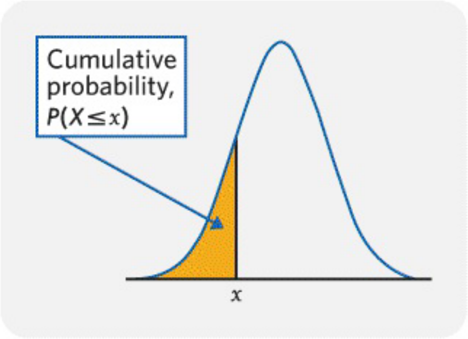
FINDING NORMAL PROBABILITIES

Areas under a Normal curve represent proportions (frequencies) of observations from that Normal distribution. Therefore, they also represent probabilities of randomly selecting an individual from that Normal distribution. There is no simple way, mathematically, to obtain areas under a Normal curve because of its complexity. Normal computations require either using software that calculates areas or looking up areas in a published table.

Some applets and software functions calculate the area under a Normal distribution for any interval on the horizontal axis. You would simply need to specify the boundaries of the interval, along with the mean and standard deviation of the specific Normal curve. However, published tables and many software functions provide just one kind of area, *cumulative probabilities*.

CUMULATIVE PROBABILITIES

The **cumulative probability** for a value xx in a distribution is the proportion of observations in the distribution that lie at or below x.x.



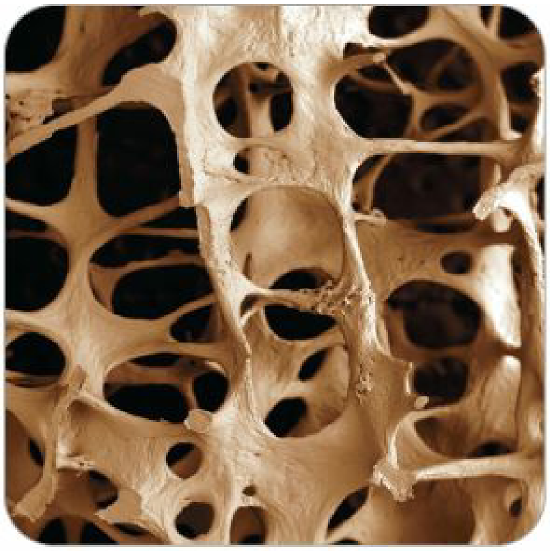
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

EXAMPLE 11.4 Defining osteoporosis

Osteoporosis is a condition in which the bones become brittle to the breaking point due to loss of minerals. To diagnose osteoporosis, bone mineral density (BMD) is measured and typically reported as a unitless score, which is designed so that the mean BMD score of healthy young adults of the same sex is zero. The World Health Organization’s defining criterion for osteoporosis is a BMD score *below* −2.5.

Women in their 70s tend to have much lower bone density than young adult women. In this population of older women, BMD scores are approximately Normally distributed, with a mean of −2 and a standard deviation of 1.[4] We want to find the probability that a randomly selected woman in her 70s would have osteoporosis.

The proportion of women in their 70s with a BMD score of −2.5 or less is the area under the curve to the left of −2.5, as shown in Figure 11.4. This is the cumulative probability up to −2.5 for the N(-2,1)$N(-2, 1)$ density curve. Figure 11.5 gives the output from a graphing calculator and three software functions. All give the probability of a BMD score below −2.5 in this population as 0.3085. There is a 30.85% chance that a randomly selected woman in her 70s would have osteoporosis.



Alan Boyde/Getty Images

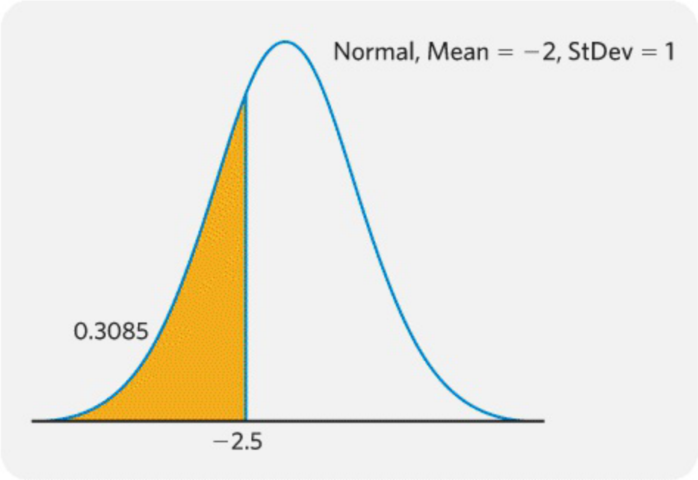
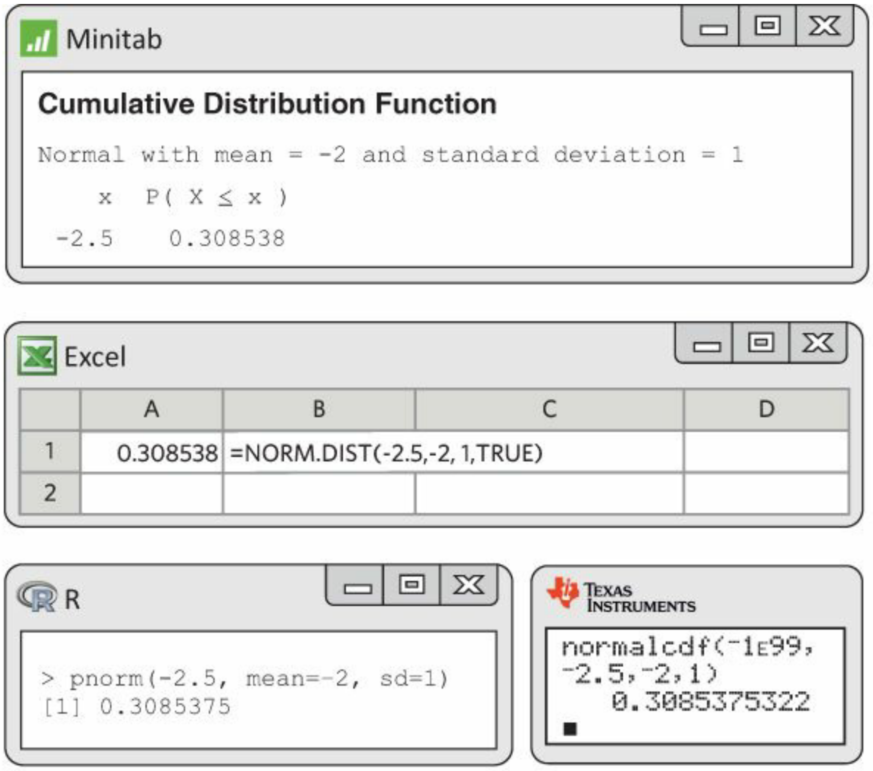


Figure 11.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.4 Minitab graph displaying the area to the left of −2.5 under the N(-2,1)

$$N(-2,1)$$

density curve. This area, 0.3085, represents the probability that a randomly selected woman in her 70s would have osteoporosis.



Cumulative Distribution Function

```
Normal with mean = -2 and standard deviation = 1

    x   P( X ≤ x )
 -2.5     0.308538
```

	A	B	C	D
1	0.308538	=NORM.DIST(-2.5,-2, 1,TRUE)		
2				

```
> pnorm(-2.5, mean=-2, sd=1)
[1] 0.3085375
```

```
normalcdf(⁻1E99,
⁻2.5,⁻2,1)
    0.3085375322
■
```

Figure 11.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.5 Software output for the cumulative probability up to −2.5 for the N(-2,1) $N(-2,1)$ density curve.

The *P* in the Minitab output stands for “probability,” but we can also read it as “proportion of the observations.” Notice that Minitab refers to P(X≤x)$P(X \le x)$ in its output. *There is no area under a smooth curve and exactly over the point* −2.5 (because this would be an “interval” with zero width). Consequently, the area to the left of −2.5 (the proportion of older women with BMD <-2.5< -2.5) is the same as the area at or to the left of this point (the proportion of older women with

BMD ≤-2.5≤ -2.5).

Minitab asks for the mean and standard deviation of the Normal distribution in a pop-up menu and displays these values in the output. Others require them within their command functions: norm.dist (x, mean, standard_dev, cumulative) for Excel, pnorm (x, mean=, sd=) for R, and normalcdf (lower, upper,μ,μ,σσ) for the TI-83.

Notice that the TI-83 normalcdf function is programmed to give the probability of any interval under a Normal curve, rather than just cumulative probabilities. Because Normal curves are defined for all real values, from negative infinity to positive infinity, the value −1E99 (or-1×1099$(\text{or} - 1 \times 10^{99}$) is used in place of negative infinity (-∞),$(-\infty)$, as it is the largest negative value that the calculator can handle.

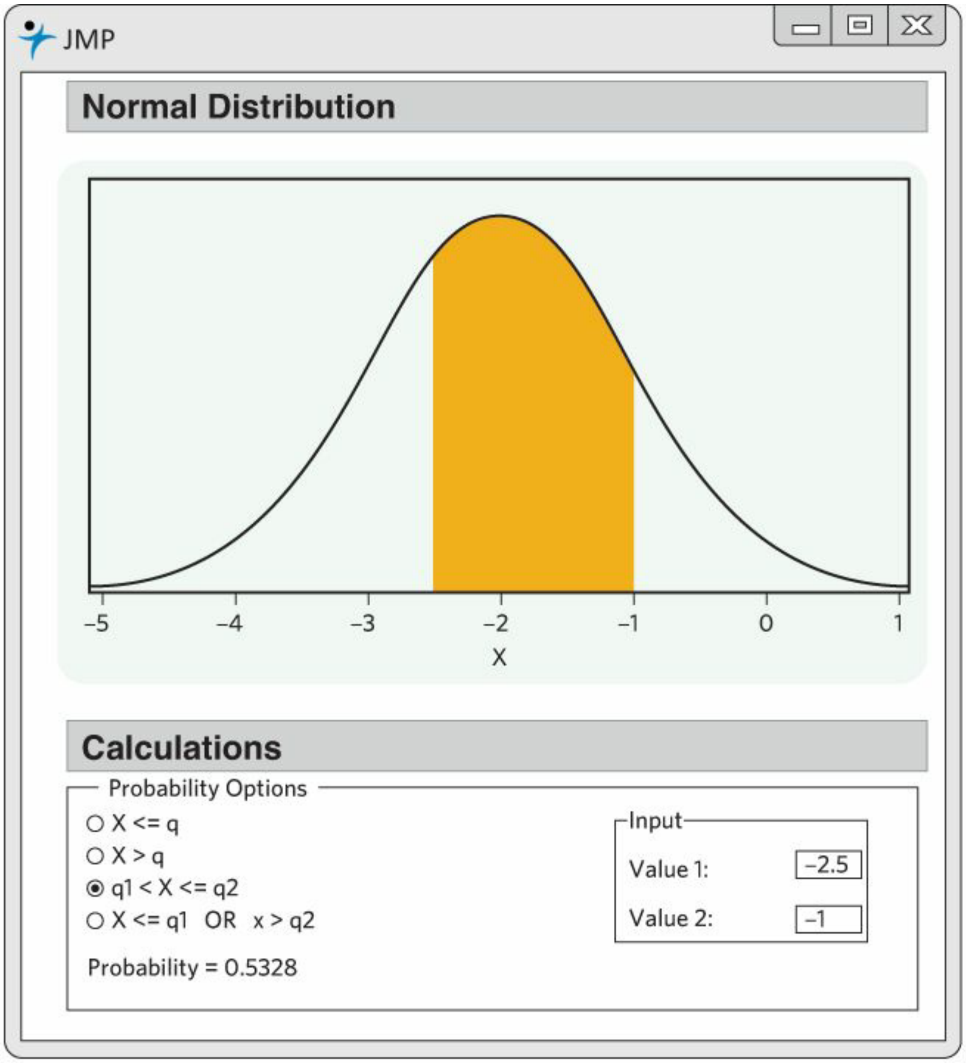
Sometimes you may need a probability other than a cumulative probability. If your technology provides only cumulative probabilities, the key is to match the area you want with the corresponding cumulative probability or the probability of its complement. *If you make a sketch of the area you want, you will almost never go wrong.*

For example, if we want to know what proportion of women in their 70s do *not* have osteoporosis, we can see from Figure 11.4 that this proportion is the complement of the shaded area. Using probability Rule 4 (page 223), we know that P(no osteoporosis)=1-P(osteoporosis)=1-0.3085=0.6915.$P(\text{no osteoporosis}) = 1 - P(\text{osteoporosis}) = 1 - 0.3085 = 0.6915.$ Therefore, 69.15% of women in their 70s do not have osteoporosis. Here is another example.

EXAMPLE 11.5 Defining osteopenia

Osteopenia is a condition defined by less than ideal bone density, corresponding to BMD scores between −2.5 and −1. Osteopenia is a greater concern for older individuals, especially postmenopausal women, in part because they tend to lose bone density more rapidly as they age. We want to find what percent of women in their 70s have osteopenia.

The proportion of women in their 70s with a BMD score between −2.5 and −1 can be obtained directly with some tools. Figure 11.6 shows the JMP software output, which gives both a numerical value of 0.5328 and the corresponding shaded area. Likewise, using the TI-83 function normalcdf(−2.5,−1,−2,1) returns the value 0.5328. This is the area under the N(-2,1)$N(-2,1)$ density curve for BMD scores between −2.5 and −1.



Probability Options

○ X <= q
○ X > q
● q1 < X <= q2
○ X <= q1 OR x > q2

Probability = 0.5328

Input

Value 1: −2.5

Value 2: −1

Figure 11.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.6 JMP output for the area between −2.5 and −1 under the N(-2,1) $N(-2,1)$ distribution. The shaded area is 0.5328 and represents the proportion of older women diagnosed with osteopenia.

Figure 11.7 illustrates graphically the approach that we must take when the tool used for Normal computations provides only cumulative probabilities. The area under the N(-2,1)$N(-2, 1)$ density curve for values between −2.5 and −1 is equal to the cumulative area up to −1 minus the cumulative area up to −2.5. In probability notation, this would be equivalent to stating

P(-2.5<X≤-1)=P(X≤-1)-P(X≤-2.5)0.5328=0.8413-0.3085

$$\begin{aligned} P(-2.5 < X \leq -1) &= P(X \leq -1) - P(X \leq -2.5) \\ 0.5328 &= 0.8413 - 0.3085 \end{aligned}$$

Therefore, 53.28% of women in their 70s have osteopenia.

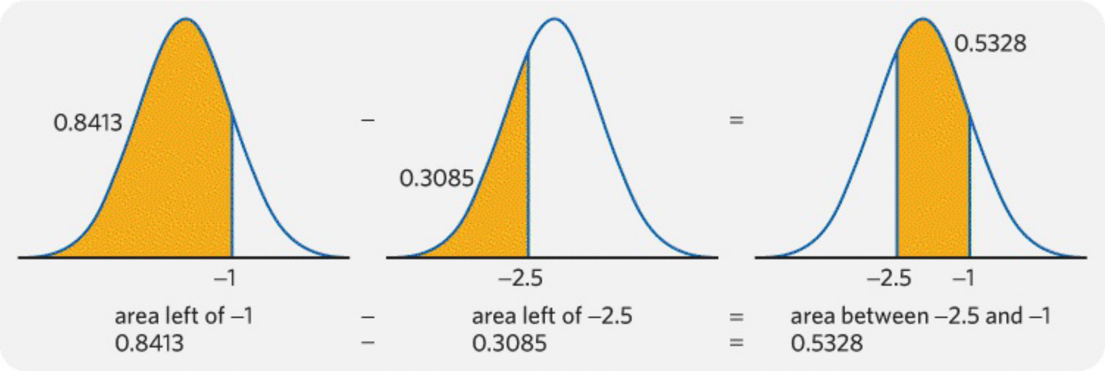


Figure 11.7

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.7 Obtaining the probability of osteopenia in older women (area between −2.5 and −1) using cumulative probabilities.

Examples 11.4 and 11.5 illustrate how a continuous random variable can be used to define a disease or a medical condition based on a chosen threshold. It is important to keep in mind, however, that there is no fundamental physiological difference between the BMD scores −2.6 and −2.4, even though they are on opposite sides of the threshold used to define osteoporosis. To some extent, the exact threshold is an arbitrary value whose purpose is to facilitate diagnosis and reporting.

APPLY YOUR KNOWLEDGE

11.7 Osteopenia, continued. Osteopenia, described in Example 11.5, is a range of suboptimal bone densities rather than an actual disease. It is somewhat controversial because the term sounds like a disease, and because the −1 threshold is not very far below the mean for healthy young adults. In fact, the risk of fracture increases gradually with more negative BMD scores.[5] Use the technology of your choice to find the following proportions in the population of women in their 70s, as in Example 11.5.

a. Women with BMD scores between −1 and −1.5
b. Women with BMD scores between −1.5 and −2
c. Women with BMD scores between −2 and −2.5

11.8 Finding a proportion or a probability. Use the technology of your choice to find the proportion of observations from a standard Normal distribution, N(0,1),$N(0, 1)$, that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

a. z<2.85$z < 2.85$
b. z>2.85$z > 2.85$
c. z>-1.66$z > -1.66$
d. -1.66<z<2.85$-1.66 < z < 2.85$
e. Explain in your own words why these proportions can also be considered probabilities.

FINDING PERCENTILES

Examples 11.4 and 11.5 illustrate how to find the proportion of the observations that satisfies some condition, such as “BDM score below -2.5.” Sometimes, however, we may want to find the *value* with a given proportion of the observations above or below it.

Earlier, we performed this task when we stated the 68–95–99.7 rule. For instance, we saw that for any Normal distribution with mean μ and standard deviation σ,

$$P(\mu-\sigma<X<\mu+\sigma)\approx 0.68$$

$$P(\mu-\sigma<X<\mu+\sigma)\approx 0.68$$

Because Normal distributions are symmetric, it follows that

$$P(X<\mu-\sigma)\approx 0.16 \text{ and} P(X<\mu+\sigma)\approx 0.84$$

$$\begin{aligned} P(X<\mu-\sigma) &\approx 0.16 \text{ and} \\ P(X<\mu+\sigma) &\approx 0.84 \end{aligned}$$

That is, in the $N(\mu,\sigma)$ distribution, approximately 16% of all observations are below the value $\mu-\sigma$, and approximately 84% of all observations are below the value $\mu+\sigma$. We say that the values $\mu-\sigma$ and $\mu+\sigma$ are, respectively, roughly the *16th and 84th percentiles* of the $N(\mu,\sigma)$ distribution.

PERCENTILE

The value such that some percent *p* of the observations in a distribution lie below it is called the ***p*th percentile.**

Note that, because Normal distributions describe continuous random variables, the value of a percentile is the same whether we define “below” strictly (excluding this specific value) or not (including it).

EXAMPLE 11.6 The third quartile of hatching weights

Glow Images/SuperStock

The hatching weights of commercial chickens can be modeled accurately using a Normal distribution with mean μ=45 $\mu = 45$ grams (g) and standard deviation σ=4 g. $\sigma = 4$ g.[6] What is the third quartile of the distribution of hatching weights?

We want to find the value x x representing the third quartile of the N(45,4) $N(45, 4)$ distribution of hatching weights. Recall from Chapter 2 that the third quartile is the value with 75% of the distribution *below* it. As shown in Figure 11.8, x x is the 75th percentile. Most statistical software will give you x x when you plug in mean 45, standard deviation 4, and cumulative probability 0.75.

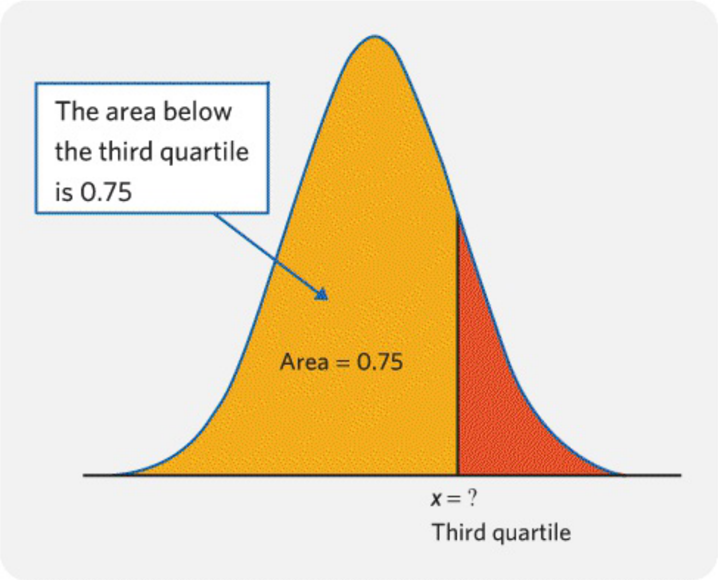
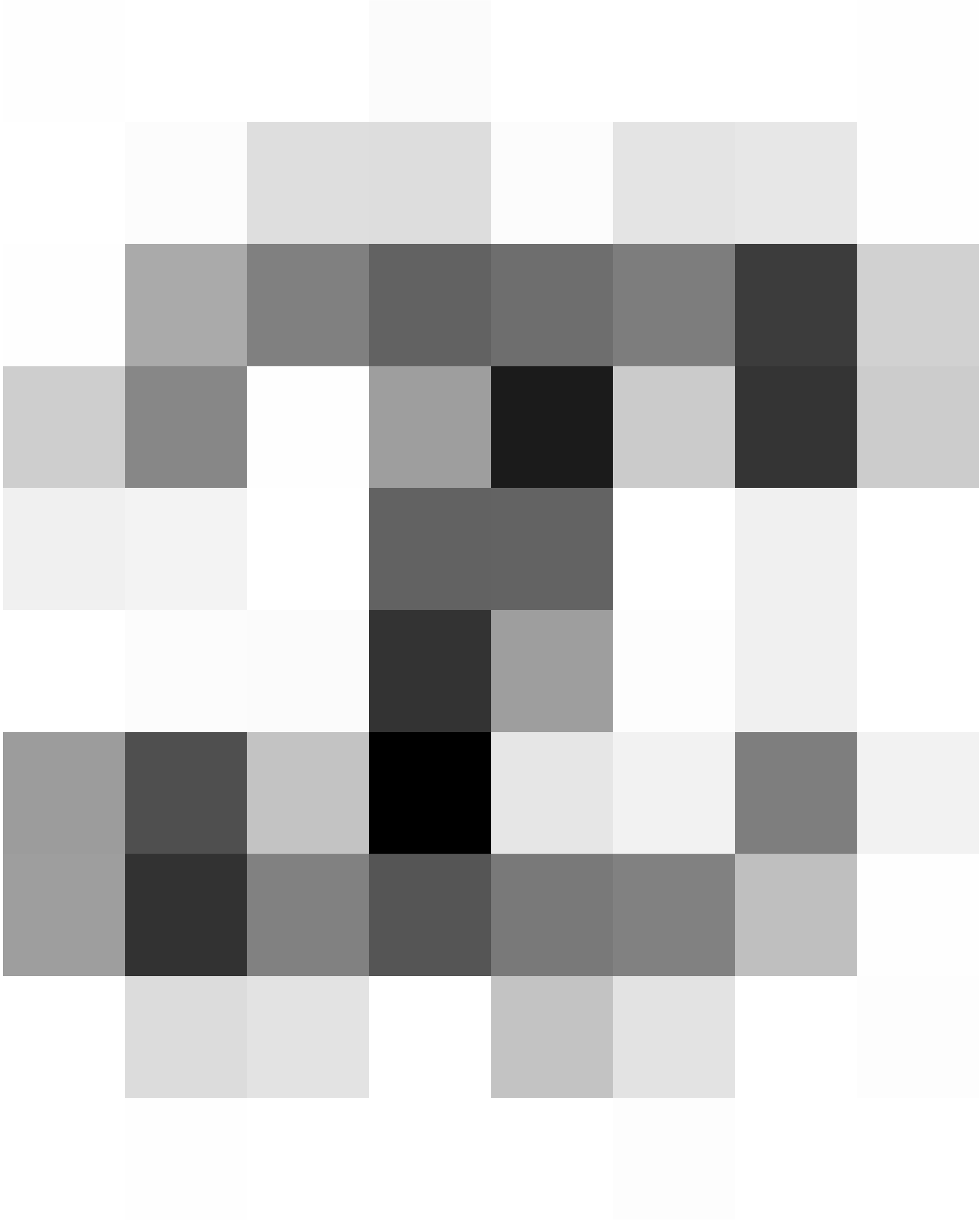


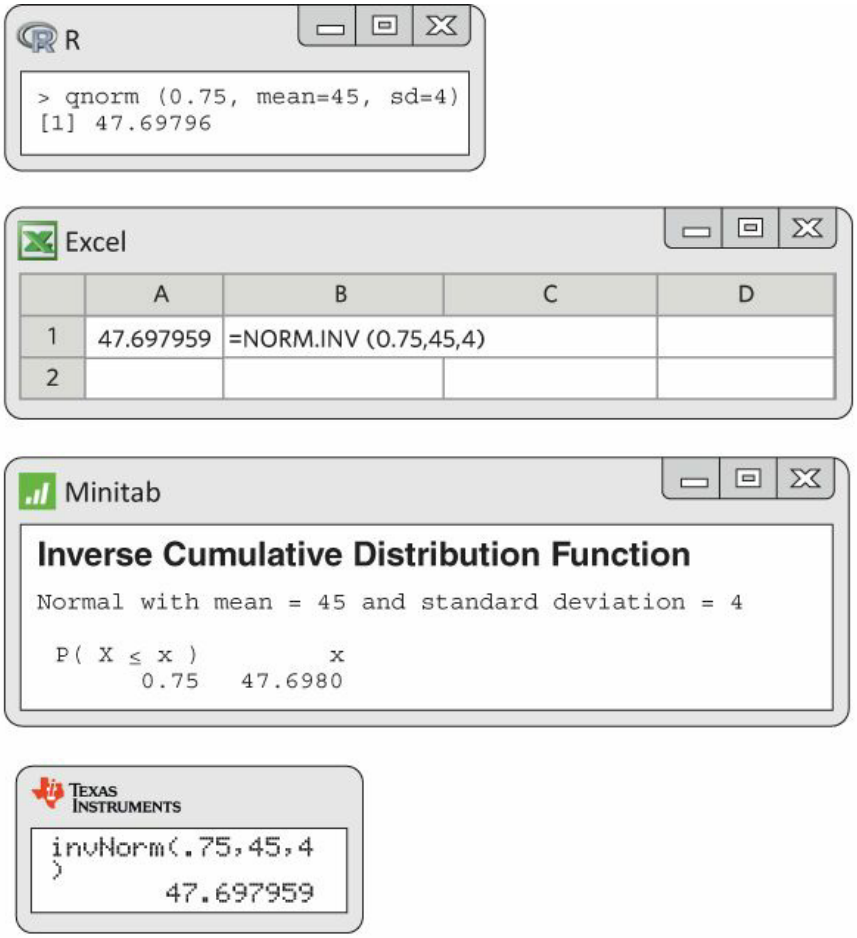
Figure 11.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.8 Locating the third quartile of a Normal curve. The third quartile x



is the 75th percentile.

Figure 11.9 displays the output from a graphing calculator and three software programs. All give approximately x=47.698.$x = 47.698$. Thus, the third quartile of hatching weights of commercial chickens is about 47.7 g. That is, 75% of commercial chickens weigh less than 47.7 g when they hatch, and 25% weigh more than 47.7 g.



```
> qnorm (0.75, mean=45, sd=4)
[1] 47.69796
```

	A	B	C	D
1	47.697959	=NORM.INV (0.75,45,4)		
2				

Inverse Cumulative Distribution Function

```
Normal with mean = 45 and standard deviation = 4

 P( X ≤ x )          x
       0.75    47.6980
```

```
invNorm(.75,45,4
)
           47.697959
```

Figure 11.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 11.9 Software output for the third quartile of the $N(45,4)$

$N(45,4)$

density curve used to model the distribution of hatching weights.

Statistical software will readily find percentiles under any Normal distribution. However, you need to pay close attention to the wording of the question. The percent stated may correspond to the area below the percentile you are looking for, or it may correspond to its complement.

The question may also ask for a single value (a percentile) or for a range of values (such as the lowest 5% or the highest 20%).

EXAMPLE 11.7 The top 10% of pregnancy lengths

The lengths of human pregnancies from conception to birth (measured in days) follow approximately the N(266,16)$N(266, 16)$ distribution. How long are the *longest* 10% of pregnancies?

Fuse/Getty Images

The questions asks for a range of values: the longest 10% of pregnancy lengths. Normal curves are mathematical functions with no boundaries. Thus, all we need to compute is the lower end of the top 10% range, the pregnancy length xx with area 0.1 *above x* (to its right) under the N(266,16)$N(266, 16)$ curve.

As we saw in Example 11.6, most software give a value xx when we plug in the distribution mean and standard deviation, and the cumulative probability up to x.x. When solving a Normal problem, it’s always a good idea to sketch the Normal curve and highlight the information provided and the desired answer. Figure 11.10 does that: It shows clearly that the pregnancy length xx with area 0.1 *above x* (to its right) is the same as the pregnancy length xx with area 0.9 *below x* (to its left). Because the N(266,16)$N(266, 16)$ distribution describes the lengths of *all* pregnancies, the value xx separates the shortest 90% of pregnancies from the longest 10%. The value xx we are looking for is therefore the 90th percentile.

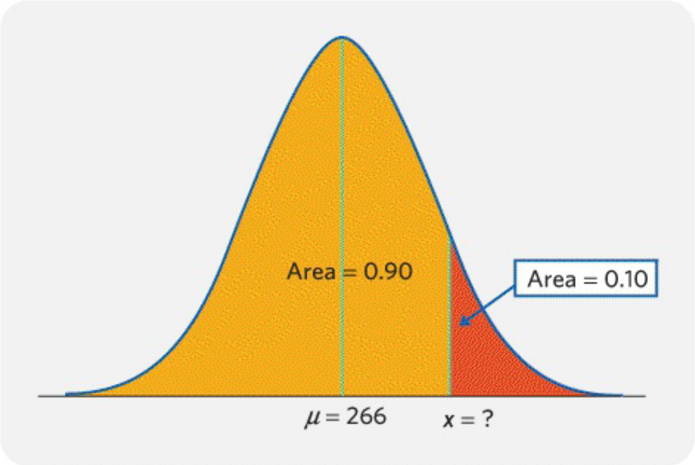
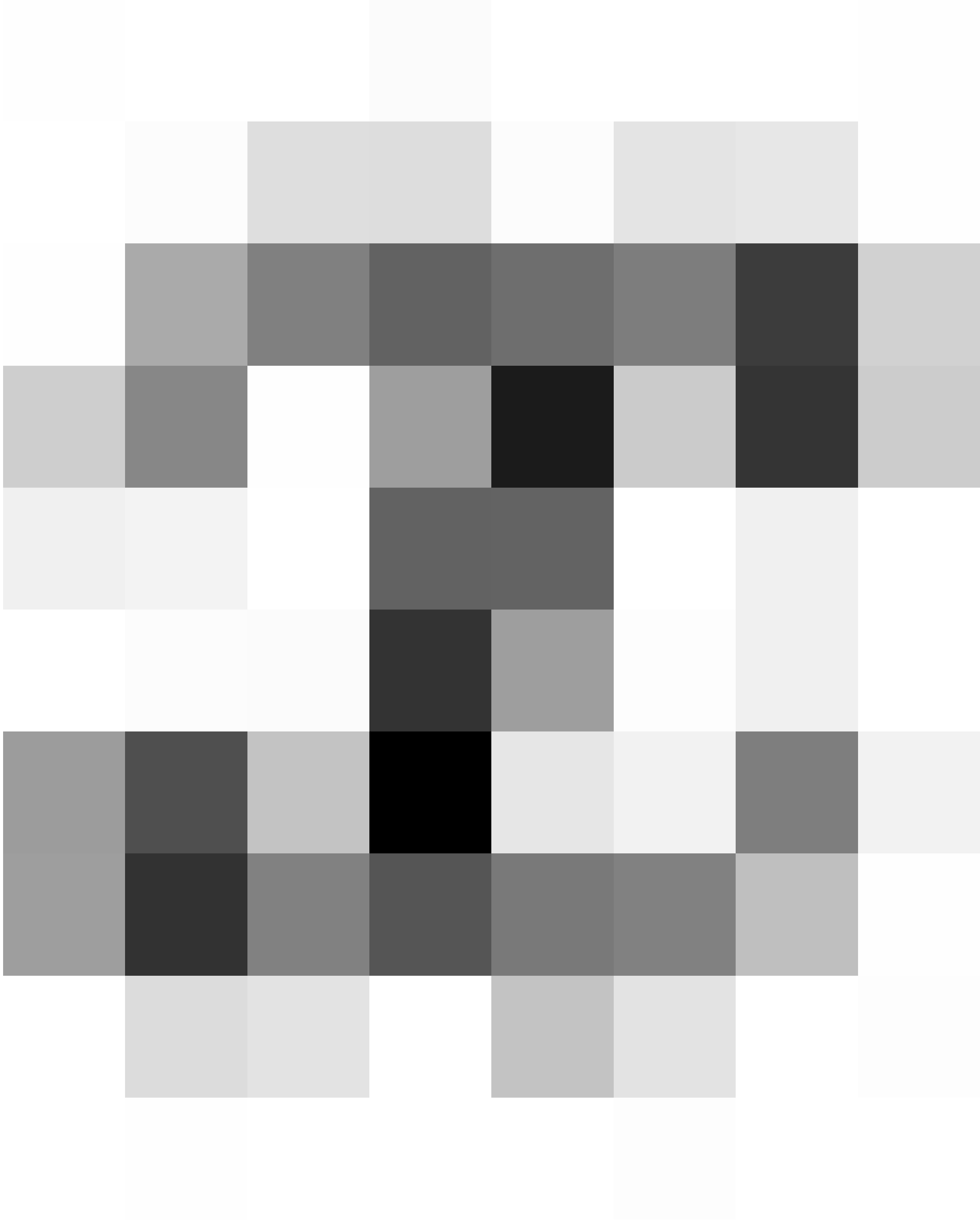


Figure 11.10
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 11.10 Finding the 10% longest pregnancies. The value x



we need to find is the 90th percentile of pregnancy lengths.

We can now use software or a calculator to obtain the value x.x. Entering either invNorm(0.9,266,16) in the TI-83 calculator or qnorm(0.9,mean=266,sd=16)$\text{qnorm}(0.9, \text{mean} = 266, \text{sd} = 16)$ in R gives approximately x=286.505.$x = 286.505.$

So, how long are the longest 10% of pregnancies? They are pregnancies lasting 286.5 days or longer. Since we typically round days

> to whole numbers, we can conclude that pregnancies longer than 286 days (287 days or longer) are in the top 10% of pregnancy lengths.

Both Examples 11.6 and 11.7 required that we obtain one numerical value using technology. However, the answer for Example 11.6 was a singular value (the third quartile of hatching weights in commercial chickens), whereas the answer for Example 11.7 was a range of values (the 10% longest human pregnancies). Statistics is the science of data in context. *Always make sure that your answer makes sense in the context of the question asked.* Your answer should be expressed as a sentence using the wording of the problem and include units when applicable.

APPLY YOUR KNOWLEDGE

11.9 Bladder volumes, continued. Exercise 11.1 described the distribution of bladder volumes among adult men and adult women. Use the technology of your choice to answer the following questions. Write your answer in a sentence in context.

a. How large are the largest 10% of bladder volumes among men?
b. How large are the largest 5% of bladder volumes among women?
c. Obtain and compare the first quartile of bladder volumes for the population of adult men and for the population of adult women.

11.10 IQ test scores. Scores on the Wechsler Adult Intelligence Scale are approximately Normally distributed, with $\mu=100$$\mu = 100$ and $\sigma=15$.$\sigma = 15$. Use the technology of your choice to answer the following questions. Write your answer in a sentence in context.

a. What scores fall in the lowest 25% of the distribution?
b. How high a score is needed to be in the highest 5%?

USING THE STANDARD NORMAL TABLE[i]

Using technology is not always an option in the classroom, especially during exams. A published table can be used as a substitute in such situations. Such a table associates *z*-scores under the standard Normal distribution, N(0,1),$N(0,1)$, with their respective cumulative proportions.

Table B in the back of the book gives cumulative proportions (probabilities) for the standard Normal distribution. The pictures at the top of the table remind us that the entries are cumulative proportions, representing the areas under the curve to the left of a value *z*.

STATS IN YOUR WORLD

THE BELL CURVE?

Does the distribution of human intelligence follow the “bell curve” of a Normal distribution? Scores on IQ tests do roughly follow a Normal distribution, because a test score is calculated from a person’s answers in a way that is designed to produce a Normal distribution in the population. To conclude that *intelligence* follows a bell curve, we must agree that the test scores directly measure intelligence. Many psychologists do not believe there is one human characteristic that we can call “intelligence” and that can be measured by a single test score.

EXAMPLE 11.8 The standard Normal table

What proportion of observations from a standard Normal variable *z* take values less than 1.47?

Solution: To find the area to the left of 1.47, locate “1.4” in the left-hand column of Table B, then locate the remaining digit 7 as “.07” in the top row. The entry opposite 1.4 and under .07 is 0.9292. This is the cumulative proportion we seek. Figure 11.11 illustrates this area.

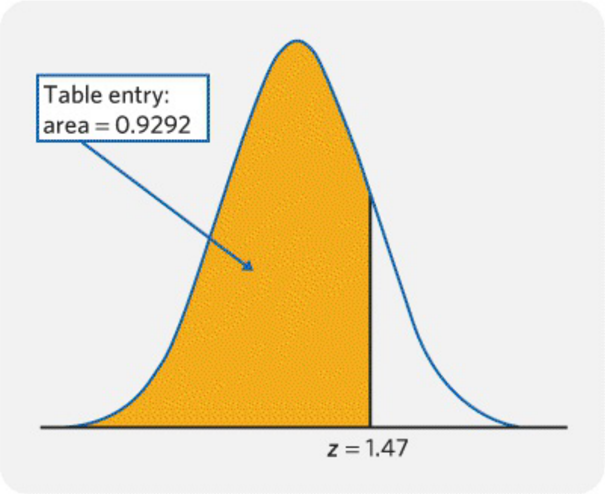


Figure 11.11
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 11.11 The area under a standard Normal curve to the left of the point z=1.47

$$z = 1.47$$

is 0.9292. Table B gives cumulative areas under the standard Normal curve.

To use just one table (like Table B) for all Normal distributions, we must first express any problem in the standard scale.

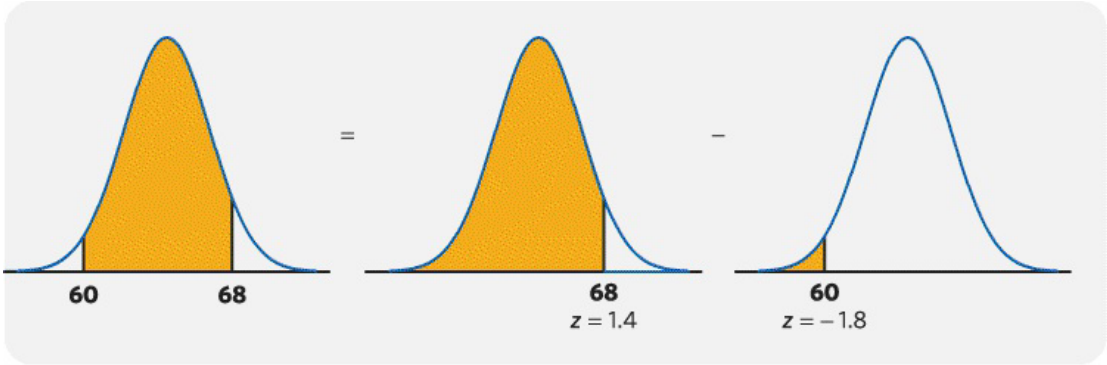
FINDING NORMAL PROBABILITIES WITH TABLE B

1. State the problem in terms of the observed variable x.x. **Draw a picture** that shows the proportion you want in terms of cumulative proportions.
2. **Standardize** xx to restate the problem in terms of a standard Normal variable *z*.
3. **Use Table B** and the fact that the total area under the curve is 1 to find the required area under the standard Normal curve.

EXAMPLE 11.9 Heights of young women: Table B

The heights of young women age 18 to 24 are approximately Normal, with μ=64.5 inches$\mu = 64.5 \text{ inches}$ and σ=2.5 inches.$\sigma = 2.5 \text{ inches}.$ What is the probability that a randomly selected woman from this age group measures between 60 and 68 inches?

STEP 1. State the problem and draw a picture. Call the height x.x. The variable xx has the N(64.5, 2.5)$N\,(64.5,\ 2.5)$ distribution. The probability that a randomly selected young woman measures between 60 and 68 inches reflects the proportion of women measuring between 60 and 68 inches in the population of these young women. What is this proportion? Here is the picture:



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

STEP 2. Standardize. Subtract the mean, then divide by the standard deviation, to turn xx into a standard Normal *z*:

60<x≤6860-64.52.5<x-64.52.5≤68-64.52.5-1.8<z≤1.4

$$\begin{aligned} 60 < x &\le 68 \\ \frac{60-64.5}{2.5} < \frac{x-64.5}{2.5} &\le \frac{68-64.5}{2.5} \\ -1.8 < z &\le 1.4 \end{aligned}$$

STEP 3. Use the table. Follow the picture (we added the *z*-scores to the picture label to help you):

area between−1.8 and 1.4=(area left of 1.4)−(area left of −1.8)=0.9192−0.0359=0.8833

$$\begin{aligned} \text{area between } -1.8 \text{ and } 1.4 &= (\text{area left of } 1.4) - (\text{area left of } -1.8) \\ &= 0.9192 - 0.0359 = 0.8833 \end{aligned}$$

This can also be written in probability terms:

P(−1.8<z≤1.4)=P(z≤1.4)-P(z≤−1.8)=0.9192-0.0359=0.8833

$$\begin{aligned} P(-1.8 < z \leq 1.4) &= P(z \leq 1.4) - P(z \leq -1.8) \\ &= 0.9192 - 0.0359 = 0.8833 \end{aligned}$$

Approximately 88% of young women measure between 60 and 68 inches. Therefore, the probability that a randomly selected young woman measures between 60 and 68 inches is about 0.88, or 88%.

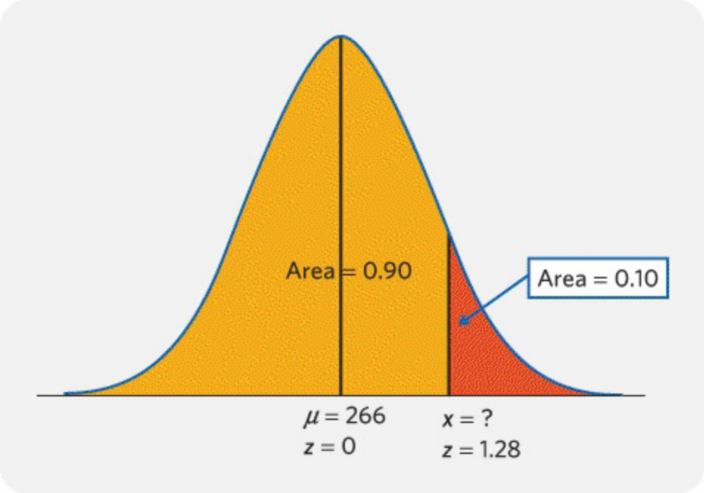
Using the R software, we get the value 0.883313 when we enter pnorm(68, mean == 64.5, sd == 2.5) −pnorm(60, mean == 64.5, sd == 2.5). We also get 0.883313 when we enter pnorm(1.4, mean == 0, sd == 1) − pnorm(−1.8, mean == 0, sd == 1). The two methods are equivalent, which is why we can compute a *z*-score and use Table B for any Normal distribution.

Table B can also be used to find a percentile, by simply using the table backward. Find the given proportion (or the nearest value to the desired proportion) in the body of the table and then read the corresponding *z* from the left column and top row. There are again three steps (with steps 2 and 3 reversed).

EXAMPLE 11.10 Find the top 10% using Table B

The lengths of human pregnancies in days from conception to birth follow approximately the N(266,16)$N(266, 16)$ distribution. How long are the longest 10% of pregnancies?

STEP 1. State the problem and draw a picture. This step is exactly as in Example 11.7. The picture is shown here.



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

STEP 2. Use the table. Look in the body of Table B for the entry closest to 0.9. It is 0.8997, the entry corresponding to z=1.28.$z = 1.28.$ So z=1.28 $z = 1.28$ is the standardized value with approximately area 0.9 to its left.

STEP 3. Unstandardize to transform *z* back to the original xx scale. We know that the standardized value of the unknown xx is z=1.28.$z = 1.28.$ So xx itself satisfies

x-26616=1.28

$$\frac{x-266}{16} = 1.28$$

Solving this equation for xx gives

x=266+(1.28)(16)=286.48

$$x = 266 + (1.28)(16) = 286.48$$

This equation should make sense: It says that xx lies 1.28 standard deviations above the mean on this particular Normal curve. That is the "unstandardized" meaning of z=1.28.$z = 1.28.$ Pregnancies longer than 286 days (287 days or longer) are among the 10% longest pregnancies.

This approach can be used to find a percentile for any Normal distribution. To find the value xx from the Normal distribution with mean μμ and standard deviation σσ corresponding to a given standard Normal value *z*, use the general formula for unstandardizing a *z*-score:

$$x = \mu + z\sigma$$

$x = \mu + z\sigma$

STATS IN YOUR WORLD

IS YOUR CHOLESTEROL TOO HIGH?

High blood cholesterol increases your risk of heart disease and stroke. The higher your total cholesterol level, the greater your risk, but there is no sharp cut-off point between a healthy level and a pathological level. Historically, total cholesterol levels greater than 240 mg/dl have been considered a serious risk factor. In a shift toward early detection and prevention, a new “elevated cholesterol” label for the 200 to 240 mg/dl range was created to identify individuals who may be at some risk of cardiovascular disease. A study estimates that this change resulted in more than 40 million new candidates for pharmacological treatment.

A printed table like Table B at the back of our book is more cumbersome than technology and will give slightly less accurate results when the calculated *z*-score is not listed in the table. (It may have more than two decimal places or it may be a value outside the range provided in the table.) In such cases, we can use the closest *z* value available in the table, though this obviously introduces some rounding error. Fortunately, Table B is detailed enough that such rounding errors are typically unimportant. If more precise computations are needed, a calculator or software function is the solution.

APPLY YOUR KNOWLEDGE

11.11 Table B. Use Table B to find the following values for the standard Normal distribution:

a. The proportion of observations such that $z > 2.85$$z > 2.85$

b. The proportion of observations such that $-1.66 < z < 2.85$

$-1.66 < z < 2.85$

c. The point *z* with 25% of the observations falling below it

d. The point *z* with 40% of the observations falling above it

11.12 Bladder volumes using Table B. Exercise 11.1 described the distribution of bladder volumes in men and in women. Use Table B to answer the following questions.

a. What proportion of bladders are between 500 and 600 ml in the adult male population?

b. What proportion of bladders are between 500 and 600 ml in the adult female population?

c. How large are the largest 25% of bladder volumes in the adult male population?

d. How large are the largest 25% of bladder volumes in the adult female population?

NORMAL QUANTILE PLOTS[ii]

The Normal distributions provide good descriptions of some distributions of real data, such as IQ tests and individual heights. In contrast, the distributions of some other common variables are skewed and, therefore, distinctly non-Normal. Examples include the survival times of cancer patients after treatment, the amount of a given toxin in the blood among a random sample of individuals, and the average size or weight of mammalian species. While experience can suggest whether a Normal distribution is plausible in a particular case, it is risky to assume that a distribution is Normal without actually inspecting the data.

A histogram or dotplot can reveal distinctly non-Normal features of a distribution, such as pronounced skew and outliers (such as the rates of pertussis for all California counties in 2014, in Figure 1.7, page 18) or clusters (such as the age of patients diagnosed with Lyme disease in Figure 1.8, page 19). If the dotplot or histogram appears roughly symmetric and unimodal (single-peaked), however, we need a more sensitive way to judge the adequacy of a Normal model (not all symmetric distributions are Normal). The most useful tool for assessing Normality is another graph, the **Normal quantile plot**.

Here is the basic idea of a Normal quantile plot. The graphs produced by software use more sophisticated versions of this idea. It is not practical to construct Normal quantile plots by hand.

1. Arrange the observed data values from smallest to largest. Record what percentile of the data each value occupies. For example, the smallest observation in a set of 20 is at the 5% point, the second smallest is at the 10% point, and so on.
2. Perform Normal distribution calculations to find the *z*-scores at these same percentiles. For example, z=-1.645$z = -1.645$ is the 5% point of the standard Normal distribution, and z=-1.282$z = -1.282$ is the 10% point.
3. Plot each data point xx against the corresponding *z*. If the data distribution is close to standard Normal, the plotted points will lie close to the 45-degree line x=z.$x = z.$ If the data distribution is close to any Normal distribution, the plotted points will lie close to some

straight line.

Any Normally distributed data set will produce a straight line on a Normal quantile plot, because the data distribution and the *z* distribution are both Normal and their relationship is thus linear. If the data are not Normally distributed, the data and the *z* distribution are unrelated, and a Normal quantile plot will not be linear.

USE OF NORMAL QUANTILE PLOTS

If the points on a **Normal quantile plot** lie close to a straight line, the plot indicates that the data are Normal. Systematic deviations from a straight line indicate a non-Normal distribution. Outliers appear as points that are far away from the overall pattern of the plot.

EXAMPLE 11.11 Shark body lengths

Mark Conlin/Getty Images

Figure 11.12 displays the distribution of the body lengths of sharks from Example 1.5 (page 13). Figure 11.12(a) is a vertical dotplot replicated from Figure 1.10(b) (page 21), which shows a symmetric distribution of shark lengths with a low outlier and a high outlier. Figure 11.12(b) is a Normal quantile plot of the same data. Most of the points lie close to a straight line, indicating that a Normal distribution describes these points quite well. The low and high outliers at 9.4 and 22.8 feet lie, respectively, below and above the line formed by the rest of the data. These are the same outliers identified in the dotplot.

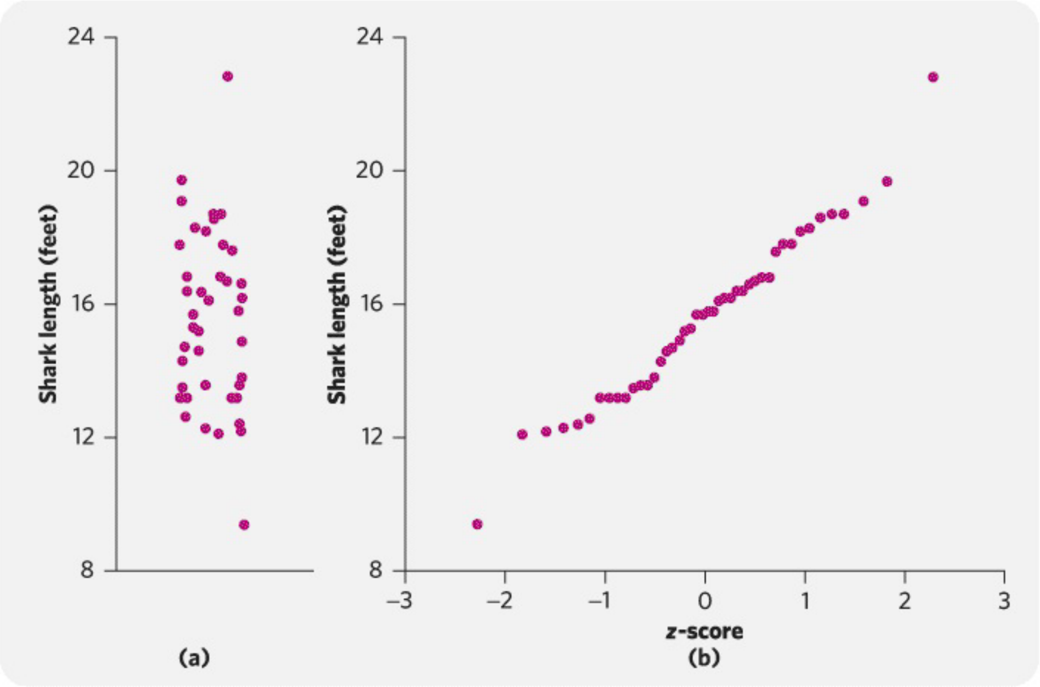


Figure 11.12

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.12 Distribution of shark lengths. (a) Vertical dotplot with jitter added for visibility, from Figure 1.10b. A low outlier and a high outlier are evident. (b) Normal quantile plot of shark lengths, which shows an approximately Normal distribution except for a low outlier and a high outlier.

EXAMPLE 11.12 Guinea pig survival times

Figure 11.13 displays the survival times in days of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment. On the left (a), a histogram shows that the data are strongly skewed to the right. The Normal quantile plot on the right (b) displays the survival data vertically against the corresponding standard Normal *z*-score plotted horizontally. The *z*-score scale extends from −3 to 3 because almost all of a standard Normal curve lies between these values.

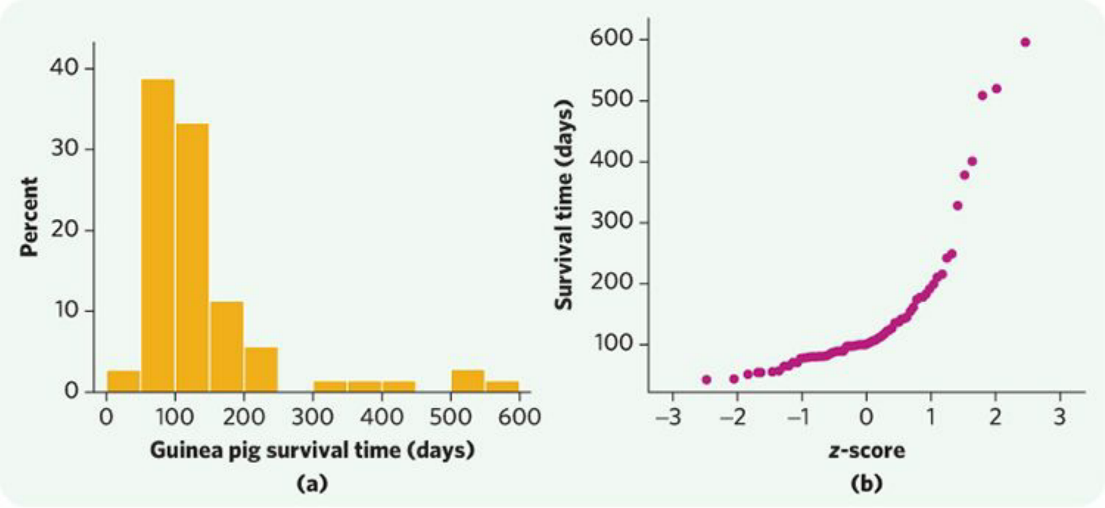


Figure 11.13
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.13 (a) Histogram and (b) Normal quantile plot of guinea pig survival times. This distribution is skewed to the right.

The points on the Normal quantile plot clearly do not follow a linear pattern. Half the data points—those with *z*-scores below zero—are survival times ranging from roughly 0 to 120 days. The other half—those with *z*-scores greater than zero—are survival times ranging from roughly 120 to 600 days. This deviation in the Normal quantile plot illustrates the right-skew.

In a right-skewed distribution, points with positive *z*-scores stretch out over the largest range of xx values. Inversely, in left-skewed distributions, points with negative *z*-scores are the ones that stretch out over the largest range of xx values.

EXAMPLE 11.13 Acid rain

Exercise 1.35 (page 33) provided data on the acidity (pH) of 105 samples of rainwater. Figure 11.14 shows (a) a histogram of these data with a Normal curve fitted over it and (b) a Normal quantile plot of the same data. Most of the points in the Normal quantile plot lie close to a straight line, indicating that a Normal distribution describes these points quite well. Notice also that the majority of the points on the Normal quantile plot have *z*-scores between −1 and 1, while few have *z*-scores close to 3 or −3, reflecting the 68–95–99.7 property of Normal distributions.

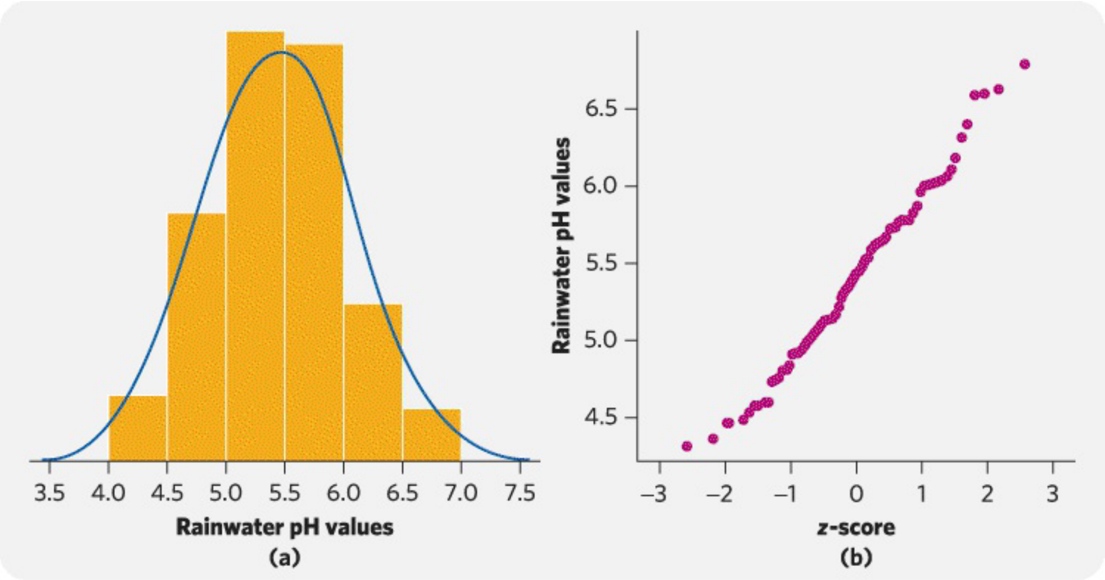


Figure 11.14

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.14 (a) Histogram with superimposed Normal curve and (b) Normal quantile plot of rainwater pH values. This distribution is roughly Normal.

As Exercise 1.35 illustrated, histograms don't settle the question of approximate Normality, because their particular shape depends on the choice of classes. The Normal quantile plot makes it clear that a Normal distribution is a good description for these data—there are only minor wiggles in a generally straight-line pattern.

As Figure 11.14 illustrates, real data almost always show some departure from the theoretical Normal model. *When you examine a Normal quantile plot, look for shapes that show clear departures from Normality. Don't overreact to minor wiggles in the plot.* When we discuss statistical methods that are based on the Normal model, we will pay attention to the sensitivity of each method to departures from Normality. Many common methods work well as long as the data are approximately Normal and outliers are not present.

CAUTION

APPLY YOUR KNOWLEDGE

11.13 Heart rate. Figure 11.15 is a Normal quantile plot of the heart rates of 200 male runners after a 6-minute treadmill run.[7] The distribution is close to Normal. How can you see this? Describe the nature of the small deviations from Normality that are visible in the plot.

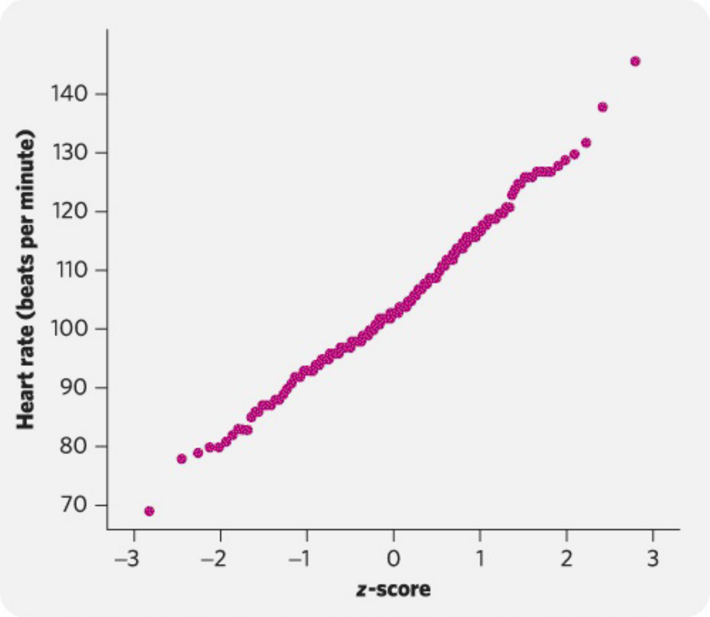


Figure 11.15

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.15 Normal quantile plot of the heart rates of 200 male runners.

11.14 Transforming data: counting seeds. Table 3.5 (page 89) gives data on the mean number of seeds produced in a year by several common tree species. (Some species appear twice because their seeds were counted in two locations.)

a. Figure 11.16(a) is a Normal quantile plot of these seed count values. In what way is the distribution of seed counts non-Normal? Is it right-skewed or left-skewed? Look carefully at Figure 11.16(a): Are there more points on the graph corresponding to small seed counts or to large seed counts?

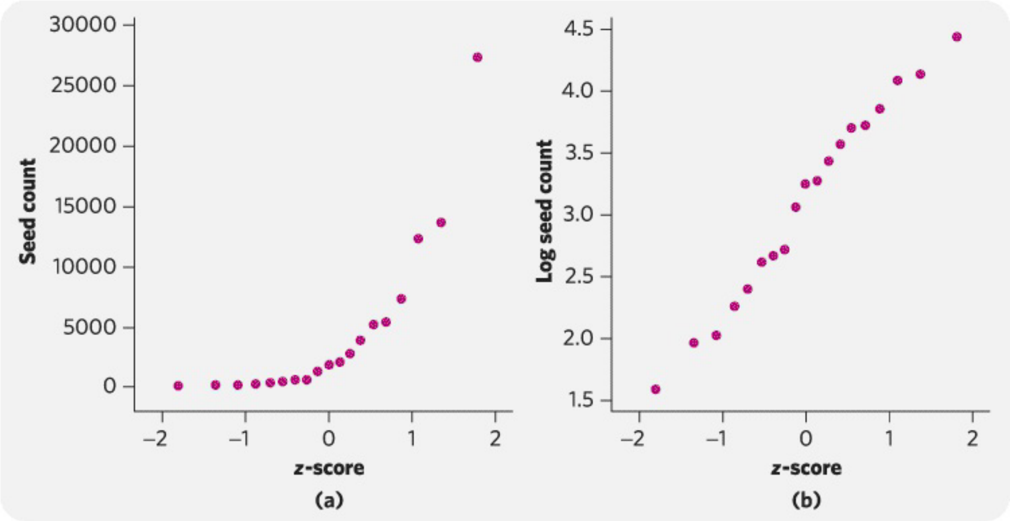


Figure 11.16

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.16 Normal quantile plots of (a) seed count for 19 tree species and (b) logarithm of seed count for these tree species.

b. Biologists know that, when dealing with sizes and counts, the logarithms of the original data are often a judicious choice of variable. Figure 11.16(b) is a Normal quantile plot of the logarithms (log) of the original seed count values. Explain why we can say that the distribution of the log of seed counts is approximately Normal. *Note: When a variable x is not Normally distributed but its logarithm log(x) is, we say that x has a lognormal distribution.*

CHAPTER 11 SUMMARY

- We can describe the overall pattern of a continuous distribution by a **density curve.** A density curve has total area 1 underneath it. An area under a density curve gives the proportion of observations that fall within a range of values.
- The **Normal distributions** are described by a special family of bell-shaped, symmetric density curves, called **Normal curves.** The mean μμ and standard deviation σσ completely specify a Normal distribution N(μ,σ).$N(\mu,\sigma)$. The mean is the center of the curve, and σσ is the distance from μμ to the change-of-curvature points on either side.
- To **standardize** any observation x,x, subtract the mean of the distribution and then divide by the standard deviation. The resulting ***z*-score**

  z=x-μσ

  $$z=\frac{x-\mu}{\sigma}$$

  says how many standard deviations xx lies away from the distribution mean.
- All Normal distributions are the same when measurements are transformed to the standardized scale. In particular, all Normal distributions satisfy the **68–95–99.7 rule,** which describes what percents of observations lie within one, two, and three standard deviations of the mean.
- If xx has the N(μ,σ)$N(\mu,\sigma)$ distribution, then the **standardized variable** z=(x-μ)/σ$z=(x-\mu)/\sigma$ has the **standard Normal distribution** N(0,1) $N(0,1)$ with mean 0 and standard deviation 1. Standardizing can help you express different Normal distributions on a common scale.
- Most statistical programs give a **cumulative proportion** that represents the area under a Normal distribution below a given value. **Table B** at the back of our book does the same after standardizing, by providing a printout of cumulative proportions for the standard Normal curve.

- Software and Table B can also be used to find a **percentile** value. The *p*th percentile is the value of a random variable such that *p* percent of the observations in the distribution lie below it.
- The adequacy of a Normal model for describing a distribution of data is best assessed by a **Normal quantile plot,** which can be created with most statistical software packages and graphing calculators. A pattern on such a plot that deviates substantially from a straight line indicates that the data are not Normal.

THIS CHAPTER IN CONTEXT

Some data sets can be shown to closely follow a Normal distribution. When this is true, the description of the data can be greatly simplified without much loss of information. We can calculate the percent of the distribution in an interval for *any* Normal distribution if we know its mean μ and standard deviation σ. This is also why the mean and standard deviation of a random variable, as first defined in Chapter 9, can be important numerical summaries.

Nevertheless, not all distributions can be well approximated by a Normal curve. When the distribution of the variable in the population is unknown, we can plot sample data using a histogram or, better yet, a Normal quantile plot to determine whether the sample data are roughly Normal. Because the distribution of data from a random sample is likely to reflect the true distribution of the population, roughly Normal sample data suggest that, in the population from which we sampled, the variable is likely to also be roughly Normal. This approach is key to checking whether the assumptions of many statistical inference procedures in Part III of this book are met.

Normal distributions are also good approximations to many kinds of chance outcomes, and we will use Normal curves to describe some sampling distributions in Chapter 13. When we can use a Normal distribution to model the sampling distribution of a statistic, we will be able to use Normal calculations like those performed in this chapter to obtain confidence intervals and *P*-values for hypothesis tests. We will examine these concepts in detail in Chapters 14 and 15.

CHECK YOUR SKILLS

11.15 Which of these variables is most likely to have a Normal distribution?

a. Lengths of newly hatched pythons
b. Weight category (underweight, normal weight, overweight, obese) of adults
c. Number of heart attacks experienced by men before the age of 65

11.16 To completely specify the shape of a Normal distribution, you must give

a. the mean and the standard deviation.
b. the five-number summary.
c. the mean and the median.

The respiratory rate (in breaths per minute) in newborns varies according to a distribution that is approximately Normal, with mean 50 and standard deviation 5. Exercises 11.17 to 11.19 use this information.[8]

11.17 Approximately 95% of all newborn respiratory rates (in breaths per minute) are between

a. 45 and 55.
b. 40 and 60.
c. 35 and 65.

11.18 What is the approximate probability that a randomly chosen newborn has a respiratory rate of 55 breaths per minute or more?

a. 0.16
b. 0.68
c. 0.84

11.19 What is the approximate probability that a randomly chosen newborn has a respiratory rate between 40 and 55 breaths per minute?

a. 0.498

b. 0.815
c. 0.997

11.20 What is the approximate proportion of observations from a standard Normal distribution that take values less than 1.15?

a. 0.1251
b. 0.8531
c. 0.8749

11.21 What is the approximate proportion of observations from a standard Normal distribution that take values larger than −0.75?

a. 0.2266
b. 0.7734
c. 0.8023

The common fruit fly, Drosophila melanogaster, *is the most studied organism in genetic research because it is small, easy to grow, and reproduces rapidly. The length of the thorax (where the wings and legs attach) in a population of male fruit flies is approximately Normal, with mean 0.800 millimeter (mm) and standard deviation 0.078 mm. Exercises 11.22 to 11.24 use this information.*

11.22 What proportion of male fruit flies have thorax length between 0.9 and 1 mm?

a. 0.0254
b. 0.0947
c. 0.5398

11.23 Approximately 16% of male fruit flies have thorax length shorter than

a. 0.644 mm.
b. 0.722 mm.
c. 0.878 mm.

11.24 What is the 30th percentile of thorax lengths among male fruit fly?

a. 0.759 mm

b. 0.841 mm
c. 30%

CHAPTER 11 EXERCISES

For all Normal computations, use the approach your instructor recommends for your class. You can also double-check your work with a variety of freely available online applets.

11.25 IQ test scores. The Wechsler Adult Intelligence Scale (WAIS) is the most commonly used “IQ test.” The scale of scores is set separately for each age group and is approximately Normal with mean 100 and standard deviation 15. Use the 68–95–99.7 rule to answer the following questions.

a. What percent of people have WAIS scores above 100? Above 145? Below 85?

b. The organization MENSA, which calls itself “the high IQ society,” requires an IQ score of 130 or higher for membership. What percent of adults would qualify for membership?

c. People with WAIS scores below 70 are considered mentally retarded when, for example, applying for Social Security disability benefits. About what percent of adults are mentally retarded according to this criterion?

11.26 Pregnancy length in horses. Bigger mammals tend to carry their young longer before giving birth. The length of horse pregnancies from conception to birth varies according to a roughly Normal distribution, with mean 336 days and standard deviation 3 days. Use the 68–95–99.7 rule to answer the following questions.

a. Almost all (99.7%) horse pregnancies fall within what range of lengths?

b. What percent of horse pregnancies are longer than 339 days?

11.27 The standard Normal curve. Find the proportion of observations from a standard Normal distribution that fall within each of the following regions. In each case, sketch a standard Normal curve and shade the area representing the region.

a. $z \leq -2.25$$z \leq -2.25$
b. $z \geq -2.25$$z \geq -2.25$
c. $z > 1.77$$z > 1.77$
d. $-2.25 < z < 1.77$$-2.25 < z < 1.77$

11.28 The standard Normal curve. Find the proportion of observations from a standard Normal distribution that satisfy each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

a. $z < 1.85$$z < 1.85$
b. $z > 1.85$$z > 1.85$
c. $z > -0.66$$z > -0.66$
d. $-0.66 < z < 1.85$$-0.66 < z < 1.85$

11.29 The standard Normal curve. Find the standardized value *z* that satisfies each of the following conditions (report the value of *z* that comes closest to satisfying the condition).

a. The probability is 0.8 that a randomly selected observation falls below *z*.
b. The probability is 0.35 that a randomly selected observation falls above *z*.

11.30 The standard Normal curve. Find the *z*-score that satisfies each of the following conditions (report the value of *z* that comes closest to satisfying the condition). In each case, sketch a standard Normal curve with your value of *z* marked on the axis.

a. 20% of the observations fall below *z*.
b. 30% of the observations fall above *z*.
c. 30% of the observations fall below *z*.

11.31 Grading on the curve. Grades in large classes are sometimes approximately Normally distributed—a fact that serves as the justification for “grading on a bell curve.” A common practice for very large classes is to give 16% of students A grades, 34% B grades, 34% C grades, and 16%

D and F grades. Assuming a Normal distribution of grades, what are the ranges of *z*-scores for each of these letter grades?

11.32 Hemoglobin. The distribution of hemoglobin in grams per deciliter of blood is approximately $N(14,1)$$N(14, 1)$ in women and approximately $N(16,1)$$N(16, 1)$ in men.[9]

a. What percent of women have more hemoglobin than the men’s mean hemoglobin?
b. What percent of men have less hemoglobin than the women’s mean hemoglobin?
c. What percent of men have more hemoglobin than the women’s mean hemoglobin?

11.33 Birth weights. The distribution of baby weights at birth is left-skewed because of premies (premature babies) who have particularly low birth weights. However, within a close range of gestation times, birth weights are approximately Normally distributed. For babies born at full term (37 to 39 completed weeks of gestation), for instance, the distribution of birth weight (in grams) is approximately $N(3350, 440)$.$N(3350, 440)$.[10] Low-birth-weight babies (weighing less than 2500 grams, or about 5 pounds 8 ounces) are at an increased risk of serious health problems. Among those, very-low-birth-weight babies (weighing less than 1500 grams, or about 3 pounds 4 ounces) have the highest risk of experiencing health problems.

a. What proportion of babies born full term are low-birth-weight babies?
b. What proportion of babies born full term are very-low-birth-weight babies?

11.34 Birth weights, continued. How much difference do a couple of weeks make for birth weight? Late-preterm babies are born with 34 to 36 completed weeks of gestation. The distribution of birth weights (in grams) for late-preterm babies is approximately $N(2750,560)$.$N(2750, 560)$.

a. What is the probability that a randomly chosen late-preterm baby would have a low birth weight (less than 2500 grams)?
b. What is the probability that a randomly chosen late-preterm baby

would have a very low birth weight (less than 1500 grams)?

c. How do these results compare with your findings in Exercise 11.33?

11.35 Arsenic in blood. Arsenic is a compound that occurs naturally in very low concentrations. Arsenic blood concentrations in healthy adults are Normally distributed, with mean $\mu=3.2$$\mu = 3.2$ micrograms per deciliter (μg/dl) and standard deviation $\sigma=1.5\mu g/dl$.$\sigma = 1.5\mu g/ dl$.[11] What is the range of arsenic blood concentrations corresponding to the middle 90% of healthy adults?

11.36 Cholesterol in middle-aged men. The blood cholesterol levels of men age 55 to 64 are approximately Normal, with mean 222 milligrams per deciliter (mg/dl) and standard deviation 37 mg/dl. What percent of these men have high cholesterol (levels above 240 mg/dl)? What percent have elevated cholesterol (between 200 and 240 mg/dl)?

11.37 Ant traffic. Researchers recorded the speed of ants on trails in their natural environment. The ants studied, *Leptogenys processionalis*, all have the same body size in their adult phase, which made it easy to measure speed in units of body lengths per second (bl/s). The researchers found that, when traffic is light and not congested, ant speeds vary roughly Normally, with mean 6.20 bl/s and standard deviation 1.58 bl/s.[12]

a. What is the probability that an ant’s speed in light traffic is faster than 5 bl/s?

b. What percent of ant speeds in light traffic are slower than 6 bl/s?

c. What are the 10th and 90th percentiles of ant speeds in light traffic?

11.38 Ant traffic, continued. The study described in Exercise 11.37 also reported that, when traffic is congested due to a high density of ants on the trail, ant speeds vary roughly Normally, with mean 4.64 bl/s and standard deviation 0.60 bl/s.

a. What is the probability that an ant’s speed in congested traffic is faster than 5 bl/s?

b. What percent of ant speeds in congested traffic are slower than 6 bl/s?

c. What are the 10th and 90th percentiles of ant speeds in congested traffic?

d. Draw the distributions of ant speeds in light traffic and in congested traffic on the same graph. Be sure to scale the horizontal axis, and include each distribution’s mean and standard deviation.

11.39 Pregnancy length in humans. The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal, with mean 266 days and standard deviation 16 days.

a. What percent of pregnancies last less than 270 days (about 9 months)?

b. What percent of pregnancies last between 240 and 270 days (roughly between 8 and 9 months)?

c. How long do the longest 20% of pregnancies last?

11.40 Pregnancy length in humans, continued. The quartiles of any distribution are the values with cumulative proportions 0.25 and 0.75.

a. What are the quartiles of the standard Normal distribution?

b. Using your numerical values from part a, write an equation that gives the quartiles of the N(μ,σ)$N(\mu,\sigma)$ distribution in terms of μμ and σ.$\sigma.$

c. The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal, with mean 266 days and standard deviation 16 days. Apply your result from part b: What are the quartiles of the distribution of lengths of human pregnancies?

11.41 Body temperature. An 1868 paper by German physician Carl Wunderlich reported, based on more than a million body temperature readings, that healthy-adult body temperatures are approximately Normal, with mean μ=98.6$\mu = 98.6$ degrees Fahrenheit (°F)$({}^\circ\mathrm{F})$ and standard deviation σ=0.6 °F.$\sigma = 0.6\ {}^\circ\mathrm{F}.$ This is still the most widely quoted result for human temperature.

a. According to this study, what is the range of body temperatures that can be found in 95% of healthy adults? (We are looking for the middle 95% of the adult population.)

b. A more recent study suggests that healthy-adult body temperatures are better described by the N(98.2,0.7)$N(98.2, 0.7)$ distribution. Based on this later study, what is the range of values representing the 95% most common body temperatures in healthy adults? Compare this range with the one you obtained in part a.

11.42 Exploring models for climate change. Can the recent noticeable increase in the number of extreme weather events, such as heat waves, be explained by global warming? We can use Normal curves to understand how various climate change scenarios would affect the probability of observing such extreme weather events.[13]

a. We start by describing historical temperature variations with a model using a Normal curve such that cold weather and hot weather occupy, respectively, the tail areas below and above two standard deviations from the historical mean. Record-cold and record-hot temperatures occupy the tail areas below and above three standard deviations from the historical mean. Draw such a Normal curve, indicating on your sketch the areas corresponding to record-cold, cold, hot, and record-hot weather events.
b. Using the same scale, draw underneath your first Normal curve another Normal curve with the same standard deviation but a somewhat higher (warmer) mean temperature (roughly one standard deviation greater than the historical mean). Using your first Normal curve as a reference, indicate on your sketch the areas corresponding to record-cold, cold, hot, and record-hot weather events in this new model. Would we observe record-cold, cold, hot, and record-hot weather events more frequently or less frequently with this model?
c. Underneath these curves created in parts a and b, draw a third Normal curve with the same mean as that of your second curve but a somewhat larger standard deviation (aim for roughly one and a quarter to one and a half times as large a standard deviation). Using your first Normal curve as a reference, indicate on your sketch the areas corresponding to record-cold, cold, hot, and record-hot weather events in this new model. Would we observe record-cold, cold, hot, and record-hot weather events more frequently or less frequently with this model?

11.43 Climate change data. In Example 1.11 (page 23) we examined changes over time in the annual global temperature, based on planet-wide recordings. To make temperatures at different locations comparable, “temperature anomalies” are computed locally by measuring each local temperature relative to its local reference temperature (the 1951–1980 average). A recent analysis looked at the distribution of all local, seasonal temperature anomalies around the world (rather than a global average). The analysis showed that, in each of several time periods, the distribution of local seasonal temperature anomalies was approximately Normal. Because temperature anomalies are computed relative to the 1951–1980 reference period, summer temperature anomalies in the northern hemisphere for this time period followed approximately the standard Normal distribution, N(0,1).$N(0, 1)$. Decades later, summer temperature anomalies in the northern hemisphere over the 2005–2015 period followed approximately the N(1.6,1.3)$N(1.6, 1.3)$ distribution.[14]

a. Draw both distributions on the same graph, indicating the mean and standard deviation of each curve on your graph.
b. In the reference period, standardized summer temperature anomalies greater than 3 were considered to be extreme heat events. Based on the Normal approximation, what percent of local summer temperature anomalies in the northern hemisphere were extreme heat events in the 1951–1980 reference period?
c. Based on the Normal approximation, what percent of local summer temperature anomalies in the northern hemisphere between 2005 and 2015 were extreme heat events?
d. Based on the recording stations at numerous worldwide locations, 14.5% of temperature anomalies in the northern hemisphere were extreme heat events between 2005 and 2015, compared with 0.1% in the reference period of 1951 to 1980. Compare these values to the ones you obtained using Normal approximations.

11.44 Climate change: models versus data. In Exercise 11.42 we considered three possible scenarios of global climate change presented in a 2001 scientific report commissioned by the United Nations. In Exercise 11.43 we examined actual changes over time in global temperatures for the northern hemisphere based on research published in 2016. Which of the three theoretical models in Exercise 11.42 is best supported by the actual temperature changes described in Exercise 11.43?

The Normal Density Curve *applet allows you to do Normal calculations quickly. It is somewhat limited by the number of pixels available for use, so it can't hit every value exactly. In Exercises 11.45 and 11.46, use the closest available values. In each case, make a sketch of the curve from the applet marked with the values you used to answer the questions.*

11.45 Where are the quartiles? How many standard deviations above and below the mean do the quartiles of any Normal distribution lie? (Use the standard Normal distribution to answer this question.)

APPLET

11.46 How accurate is 68–95–99.7? The 68–95–99.7 rule for Normal distributions is a useful approximation. To see how accurate the rule is, drag one flag across the other so that the applet shows the area under the curve between the two flags.

APPLET

a. Place the flags one standard deviation on either side of the mean. What is the area between these two values? What does the 68–95–99.7 rule say this area is?

b. Repeat part a for locations two and three standard deviations on either side of the mean. Again compare the 68–95–99.7 rule with the area given by the applet.

Exercises 11.47 to 11.51 make use of the optional material on Normal quantile plots.

11.47 Acorn sizes. Figure 11.17 shows the Normal quantile plot for the volumes of 28 acorns from the Atlantic area.[15] Examine the plot for deviations from Normality and explain your conclusions.

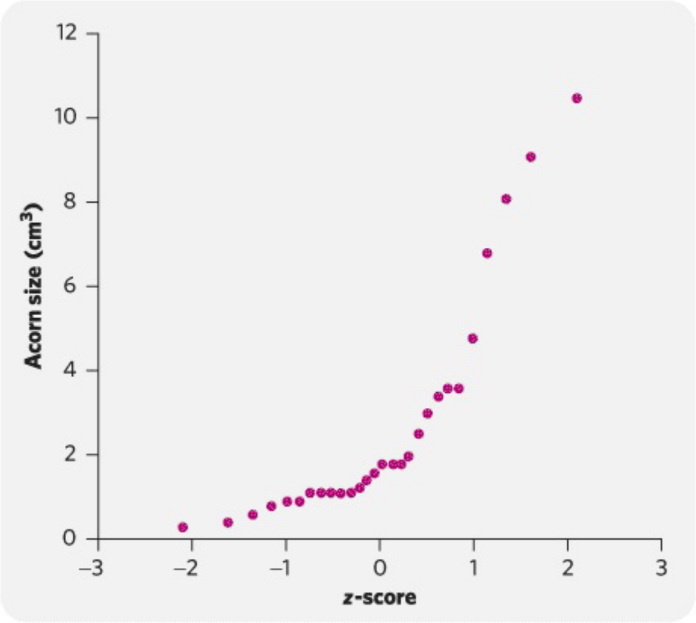


Figure 11.17
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 11.17 Normal quantile plot of the volumes of 28 acorns from the Atlantic region.

11.48 Onset of anorexia nervosa. Figure 11.18 shows the Normal quantile plot for data on the age at onset of anorexia nervosa in 691 teenage girls from Exercise 1.8. Examine the Normal quantile plot for deviations from Normality. Explain your conclusions and compare this plot with the histogram of the same data shown in Figure 1.8 (page 19).

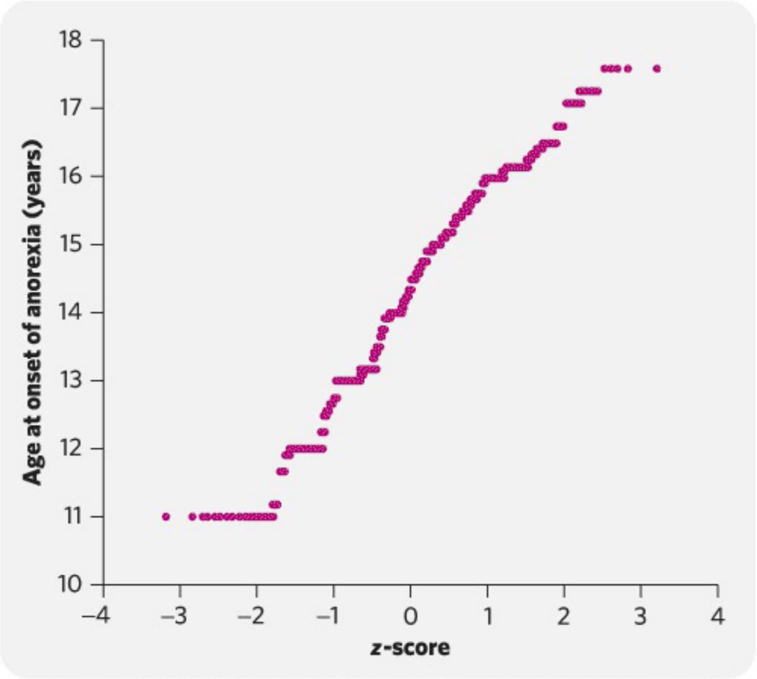


Figure 11.18

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.18 Normal quantile plot of age at onset of anorexia nervosa in 691 teenage girls.

11.49 Fruit flies. Figure 11.19 shows the Normal quantile plot for data on the thorax lengths of 49 fruit flies.[16] Examine the plot for deviations from Normality and explain your conclusions. Because of limited precision in measurements, many observations happen to have the same numerical value, which gives a steplike appearance to the Normal quantile plot. This phenomenon, called **granularity**, is not an important feature and can be ignored in your interpretation.

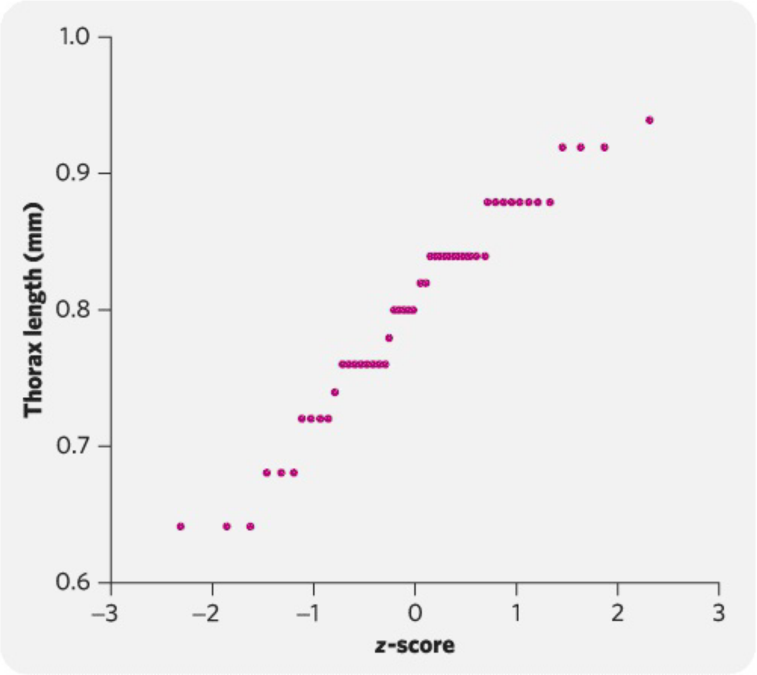


Figure 11.19

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 11.19 Normal quantile plot of the thorax lengths of 49 male fruit flies.

11.50 Logging in the rainforest. Exercise 2.37 (page 63) described a study of the effects of logging on tree counts in the Borneo rainforest.

a. Make your own Normal quantile plot of the data in Group 3 for plots of forest that were logged 8 years earlier. First, sort and rank the data from smallest to largest. Second, calculate *z*-scores for the ranks. Third, make a scatterplot of the number of trees per plot against their calculated *z*-scores.
b. Examine your Normal quantile plot for deviations from Normality and explain your conclusions.

11.51 Logging in the rainforest, continued. Exercise 2.37 (page 63) described a study of the effects of logging on tree counts in the Borneo rainforest. Use software to create Normal quantile plots for each of the three groups of forest plots. (If you completed Exercise 11.50, compare your own Normal quantile plot with the corresponding one produced by software.) Are the three distributions roughly Normal? What are the most prominent deviations from Normality that you see?

Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Note that short answers are not available for these exercises to enable their use for class work or assignments.*

11.52 Lung capacity of children (optional). The forced expiratory volume (FEV, measured in liters) is a primary indicator of lung function and corresponds to the volume of air that can forcibly be blown out in the first second after full inspiration. In Exercise 8.19 (page 209) you studied the *Large.FEV* data file containing the FEV values of a large sample of children, along with some categorical descriptor of each individual.

a. Make separate Normal quantile plots of FEV for girls (sex=0)$(\text{sex} = 0)$ and for boys (sex=1).$(\text{sex} = 1)$. What can you say from these two graphs?

b. Now use the age group variable that you created in Exercise 8.19 (*preschool* for ages 3 to 5, *elementary* for ages 6 to 10, *middle* for ages 11 to 13, and *highschool* for ages 14 and above). Make separate Normal quantile plots of FEV for each of the eight combinations of sex and age group. What can you conclude now? Explain why we must sometimes break down a data set to isolate homogeneous subgroups.

CHAPTER 12 Discrete Probability Distributions[i]

DenysKuvaiev/Depositphotos.com

IN THIS CHAPTER WE COVER...

- The binomial setting and binomial distributions
- Binomial probabilities
- Binomial mean and standard deviation
- The Normal approximation to binomial distributions
- The Poisson distributions
- Poisson probabilities

An elementary school administers eye exams to 800 students. How many students have perfect vision? A new treatment for pancreatic cancer is tried on 250 patients. How many survive for five years? You plant 10 dogwood trees. How many live through the winter? In all these situations, we want a probability model for a *count* of successful outcomes.

THE BINOMIAL SETTING AND BINOMIAL DISTRIBUTIONS

The distribution of a count depends on how the data are produced. Here is a common situation.

THE BINOMIAL SETTING

1. There are a fixed number n of observations.
2. The n observations are all **independent.** That is, knowing the result of one observation does not change the probabilities we assign to other observations.
3. Each observation falls into one of just two categories, which for convenience we call "success" and "failure."
4. The probability of a success, call it p, is the same for each observation.

Think of an obstetrician overseeing n single-birth deliveries as an example of the binomial setting. Each single-birth delivery is either a baby girl or a baby boy. Knowing the outcome of one birth doesn't change the probability of a girl on any other birth, so the births are independent. If we call a girl a "success" (this is an arbitrary choice), then p is the probability of a girl and remains the same over all the births overseen by that obstetrician. The number of girls we count is a discrete random variable X. The distribution of X is called a *binomial distribution*.

BINOMIAL DISTRIBUTION

The count X of successes in the binomial setting has the **binomial distribution** with parameters n and p. The parameter n is the number of observations, and p is the probability of a success on any one observation. The possible values of X are the whole numbers from 0 to n.

Binomial distributions are an important class of discrete probability distributions. *Pay attention to the binomial setting, because not all counts have binomial distributions.*

CAUTION

EXAMPLE 12.1 Blood donations

Blood type O is the most common blood type in America and people with this blood type can only receive blood from donors with that same type. The New York Blood Center reports that 45% of the population has blood type O.[1] Consider the next 5 blood donations from unrelated individuals. The number who have blood type O is the count X of successes in 5 independent observations with probability 0.45 of a success on each observation. So X has the binomial distribution with $n = 5$ and $p = 0.45$.

EXAMPLE 12.2 Selecting volunteer subjects

A researcher has access to 40 healthy volunteers, 20 men and 20 women. The researcher selects 10 of them at random to participate in an experiment on the effect of sport drinks taken during endurance exercise. There are 10 observations, and each is either a man or a woman. But the observations are *not* independent. If the first volunteer selected is a man, the second one is now more likely to be a woman because there are more women (20) than men (19) left in the pool of volunteers. The count X of men randomly selected in this situation does *not* have a binomial distribution.

The binomial distributions are also important in statistics when we wish to make inferences about the proportion p of “successes” in a population. Here is a typical example.

EXAMPLE 12.3 Selecting an SRS

A pharmaceutical company inspects a simple random sample (SRS) of 10 empty plastic containers from a shipment of 10,000 containers. The containers are examined for traces of benzene, a common chemical solvent but also a known human carcinogen. Suppose that (unknown to the company) 10% of the containers in the shipment contain traces of benzene. Count the number X of containers contaminated with benzene in the sample.

This is not quite a binomial setting. Just as selecting one man in Example 12.2 changes the makeup of the remaining pool of volunteers, removing one plastic container changes the proportion of contaminated containers remaining in the shipment. Thus, the probability that the second container chosen is contaminated changes when we know that the first is contaminated. Of course, removing one container from a shipment of 10,000 changes the makeup of the remaining 9999 very little. In practice, the distribution of X is very close to the binomial distribution with $n = 10$ and $p = 0.1$.

Example 12.3 shows how we can use binomial distributions in the statistical setting of selecting an SRS. When the population is much larger than the sample, a count of successes in an SRS of size n has approximately the binomial distribution, with n equal to the sample size and p equal to the proportion of successes in the population.

SAMPLING DISTRIBUTION OF A COUNT

Choose an SRS of size n from a population with proportion p of successes. When the population is much larger than the sample, the count X of successes in the sample has approximately the binomial distribution with parameters n and p.

APPLY YOUR KNOWLEDGE

In Exercises 12.1 to 12.3, the random variable X is a count. Does X have a binomial distribution? Give your reasons in each case.

12.1 Random digit dialing. When an opinion poll calls residential telephone numbers at random, only 20% of the calls reach a live person. You watch the random dialing machine make 15 calls. Xx is the number that reach a live person.

12.2 A boy or a girl? When planning a family, a couple really wants to have a boy. They decide that they will keep having children until they have their first boy. Successive births are independent, and each has about probability 0.5 of being a boy or a girl. Xx is the number of the pregnancy that succeeds in bringing a baby boy to the couple.

12.3 High school students and sex. A study of the sexual health of high school students in New York indicates that 44% have ever had sexual intercourse.[2] Xx is the number who have ever had intercourse in an SRS of 100 high school students from New York.

BINOMIAL PROBABILITIES

We can find a formula for the probability that a binomial random variable takes any value by adding probabilities for the different ways of getting exactly that many successes in nn observations. Here is an example that illustrates the idea.

EXAMPLE 12.4 Blood donations

In the population served by the New York Blood Center, each blood donor has probability 0.45 of having blood type O. In the next 5 blood donations to the center from unrelated individuals, what is the probability that exactly 2 of them have type O blood?

The count of donors with type O blood is a binomial random variable XX with n=5$n = 5$ tries and probability p=0.45$p = 0.45$ of a success on each try (as explained in Example 12.1). We want P(X=2).$P(X = 2)$.

Monkey Business Images/Shutterstock.com

Because the method doesn’t depend on the specific example, let’s use “S” for success and “F” for failure for short. Do the work in two steps.

STEP 1. Find the probability that 2 specific tries out of the 5 total—say the first and the third—give successes. This is the outcome SFSFF. Because tries are independent (the blood type of one donor doesn’t change the probability that the next unrelated donor has blood type O), the multiplication rule for independent events applies (see Chapter 10, page

243). The probability we want is

P(SFSFF)=P(S)P(F)P(S)P(F)P(F)=(0.45)(0.55)(0.45)(0.55)(0.55)=(0.45)2(0.55)3

$$\begin{aligned} P(\text{SFSFF}) &= P(\text{S})\,P(\text{F})\,P(\text{S})\,P(\text{F})\,P(\text{F}) \\ &= (0.45)\,(0.55)\,(0.45)\,(0.55)\,(0.55) \\ &= (0.45)^2\,(0.55)^3 \end{aligned}$$

STEP 2. Observe that *any one arrangement* of 2 S’s and 3 F’s has this same probability. This is true because we multiply together 0.45 twice and 0.55 three times whenever we have 2 S’s and 3 F’s. The probability that X=2$X = 2$ is the probability of getting 2 S’s and 3 F’s in any arrangement whatsoever. Here are all the possible arrangements:

SSFFF SFSFF SFFSF SFFFS FSSFF
FSFSF FSFFS FFSSF FFSFS FFFSS

There are 10 of them, all with the same probability. The overall probability of 2 successes in 5 trials is therefore

P(X=2)=10(0.45)2(0.55)3=0.3369

$$P(X = 2) = 10(0.45)^2(0.55)^3 = 0.3369$$

What if we were interested in the next 15 blood donations to the center from unrelated individuals? The pattern of calculation shown in Example 12.4 works for any binomial probability. However, it would be cumbersome to have to do the counting by actually listing all the arrangements. We would rather have a set formula. For that, we use the following fact.

BINOMIAL COEFFICIENT

The number of ways of arranging kk successes among nn observations is given by the **binomial coefficient**

nk=n!k!(n-k)!

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for k=0,1,2,…,n.$k = 0, 1, 2, \ldots, n.$

The formula for binomial coefficients uses the **factorial** notation. For any positive whole number n, its factorial $n!$ is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

Also, $0! = 1$.

The larger of the two factorials in the denominator of a binomial coefficient will cancel much of the $n!$ in the numerator. For example, the binomial coefficient we need for Example 12.4 is

$$\begin{aligned}\binom{5}{2} &= \frac{5!}{2!3!}\\ &= \frac{(5)(4)(3)(2)(1)}{(2)(1)\times(3)(2)(1)}\\ &= \frac{(5)(4)}{(2)(1)} = \frac{20}{2} = 10\end{aligned}$$

The notation $\binom{n}{k}$ *is not related to the fraction* $\frac{n}{k}$. A helpful way to remember its meaning is to read it as "binomial coefficient n choose k." Binomial coefficients have many uses, but we are interested in them only as an aid to finding binomial probabilities. The binomial coefficient $\binom{n}{k}$ counts the number of different ways in which k successes can be arranged among n observations. The binomial probability $P(X = k)$ is this count multiplied by the probability of any one specific arrangement of the *k* successes. Here is the result we seek.

BINOMIAL PROBABILITY

If X has the binomial distribution with n observations and probability p of success on each observation, the possible values of X are $0, 1, 2, \ldots, n$. If k is any one of these values,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

EXAMPLE 12.5 Inspecting pharmaceutical containers

The number X of containers contaminated with benzene in Example 12.3 has approximately the binomial distribution, with $n = 10$ and $p = 0.1$.

According to the addition rule for disjoint events, the probability that the sample contains no more than 1 contaminated container is

P(X≤1)=P(X=0)+P(X=1)=100(0.1)0(0.9)10+101(0.1)1(0.9)9=10!0!10!(1)(0.3487)+10!1!9!(0.1)(0.3874)=(1)(1)(0.3487)+(10)(0.1)(0.3874)=0.3487+0.3874=0.7361

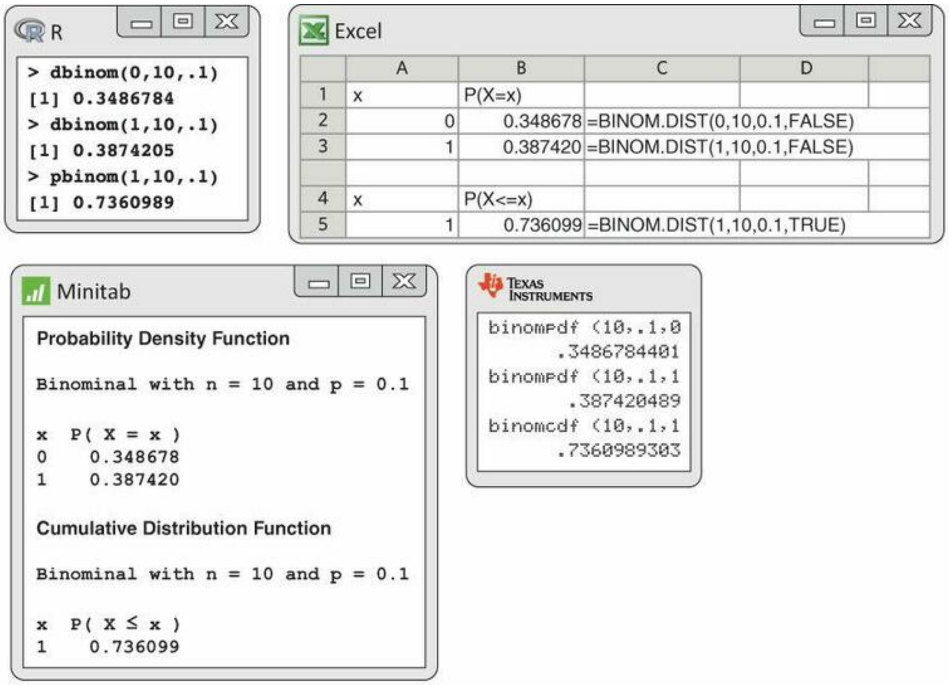
$$\begin{aligned}
P(X \le 1) &= P(X = 0) + P(X = 1) \\
&= \binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 \\
&= \frac{10!}{0!10!}(1)(0.3487) + \frac{10!}{1!9!}(0.1)(0.3874) \\
&= (1)(1)(0.3487) + (10)(0.1)(0.3874) \\
&= 0.3487 + 0.3874 = 0.7361
\end{aligned}$$

Allan Shoemake/Getty Images

This calculation uses the facts that $0! = 1$ and that $a^0 = 1$ for any number *a* other than 0. We see that about 74% of all samples will contain no more than 1 contaminated container. In fact, 35% of the samples will contain no contaminated container, so a sample of size 10 cannot be trusted to alert the pharmaceutical company to the presence of unacceptable containers in the shipment.

The binomial probability formula is useful but it is awkward to use, particularly for the probabilities of events that contain many outcomes. Historically, tables of binomial probabilities $P(X=k)$ and **cumulative probabilities** $P(X \leq k)$ have been published for selected values of n and p. The most efficient way to do binomial calculations, however, is to use technology.

Figure 12.1 shows the output for the calculation in Example 12.5 from a graphing calculator and three computer programs. We asked all four technologies to give both individual probabilities and cumulative probabilities. The names and designs of the menu items and write-in functions differ, but all four require the following information: the parameters n and p of the binomial distribution, the value of x, and an indication of whether the probability $P(X=x)$ or the cumulative probability $P(X \leq x)$ is desired.



```
> dbinom(0,10,.1)
[1] 0.3486784
> dbinom(1,10,.1)
[1] 0.3874205
> pbinom(1,10,.1)
[1] 0.7360989
```

	A	B	C	D
1	x	P(X=x)		
2	0	0.348678	=BINOM.DIST(0,10,0.1,FALSE)	
3	1	0.387420	=BINOM.DIST(1,10,0.1,FALSE)	
4	x	P(X<=x)		
5	1	0.736099	=BINOM.DIST(1,10,0.1,TRUE)	

```
Probability Density Function

Binominal with n = 10 and p = 0.1

x  P( X = x )
0    0.348678
1    0.387420

Cumulative Distribution Function

Binominal with n = 10 and p = 0.1

x  P( X ≤ x )
1    0.736099
```

```
binompdf (10,.1,0
          .3486784401
binompdf (10,.1,1
           .387420489
binomcdf (10,.1,1
          .7360989303
```

Figure 12.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 12.1 The binomial probability $P(X \leq 1)$:

$$P(X \leq 1):$$

output from a graphing calculator and three computer programs.

Note that Minitab calls P(X=x)$P(X = x)$ “Probability Density Function” and P(X≤x)$P(X \leq x)$ “Cumulative Distribution Function.” This terminology is fairly common and is reflected in the abbreviations “pdf” and “cdf” used by the TI-83. Most technology tools offering preprogrammed probability computations give a choice between a single probability and a cumulative probability. As Figure 12.1 illustrates, the distinction is often uniquely implemented. *Always confirm that you know which probability you are getting when using technology.* When in doubt, take the time to verify that you are getting P(X≤1)=P(X=1)+P(X=0) $P(X \leq 1) = P(X = 1) + P(X = 0)$ with the choice of functions offered.

For all four outputs in Figure 12.1, using the cumulative probability option was the simplest way to obtain P(X≤1),$P(X \leq 1)$, as it required only one step. We can obtain the same answer, up to roundoff error, by adding P(X=1)$P(X = 1)$ and P(X=0).$P(X = 0)$. The probability that a sample of 10 randomly selected containers would include no more than 1 contaminated container is 0.7361.

APPLY YOUR KNOWLEDGE

12.4 Blood donations, continued. In the population served by the New York Blood Center, 45% have blood type O. Consider the next 15 blood donations to the center from unrelated individuals.

a. What is the probability that exactly 3 have blood type O?
b. What is the probability that 3 or fewer have blood type O?

12.5 Antibiotic resistance. National Institute of Allergy and Infectious Diseases, National Institutes of Health Antibiotic resistance occurs when disease-causing microbes no longer respond to antibiotic drug therapy. Because such resistance is typically genetic and transferred to the next generations of microbes, it is a very serious public health problem. Gonorrhea is the second most commonly reported notifiable disease in the United States. According to the Centers for Disease Control and Prevention (CDC), 27% of gonorrhea cases tested in 2010 were resistant to at least one of the three major antibiotics commonly used to treat this sexually transmitted disease.[3] A physician treated 10 cases of gonorrhea at some point in 2010.

a. What is the distribution of cases resistant to at least one of the three major antibiotics?
b. What is the probability that exactly 1 out of the 10 cases was resistant to at least one of the three major antibiotics? What is the probability that exactly 2 out of 10 cases were resistant?
c. Find the probability that 1 or more out of the 10 cases were resistant to at least one of the three major antibiotics. (*Hint:* It is easier to first find the probability that exactly 0 of the 10 cases were resistant.)

12.6 Antibiotic resistance, continued. Antibiotic resistance is an increasing problem. The CDC reported that 38% of gonorrhea cases tested in 2014 were resistant to at least one of the three major antibiotics commonly used to treat gonorrhea.[4] Consider a physician who treated 10 cases of gonorrhea at some point in 2014.

a. What is the distribution of cases resistant to at least one of the three antibiotics?
b. What is the probability that exactly 1 out of the 10 cases was resistant to at least one of the three antibiotics? What is that probability for exactly 2 out of 10?
c. Also find the probability that 1 or more out of the 10 were resistant to all three antibiotics. (*Hint:* It is easier to first find the probability that exactly 0 of the 10 cases were resistant.)

National Institute of Allergy and Infectious Diseases, National Institutes of Health

BINOMIAL MEAN AND STANDARD DEVIATION

In Chapter 9 (page 230) we defined the mean μ and the standard deviation σ of a random variable, two important *parameters* of a probability distribution. For any particular discrete random variable, the equations on page 230 can be used to obtain the values of μ and σ. These computations can be very lengthy for binomial distributions with a large number n of observations. Fortunately, the mean and standard deviation of any binomial distribution can be computed much more easily (though the mathematical proof is beyond the scope of an introductory statistics textbook).

If a count X has the binomial distribution based on n observations with probability p of success, what is its mean μ? That is, in very many repetitions of the binomial setting, what will be the average, or *expected*, count of successes? The answer makes intuitive sense. If a psychic correctly guesses a pattern in a test 60% of the time, the mean number of correct guesses in 10 tries should be 60% of 10, or 6. In general, the mean of a binomial distribution should be $\mu = np$. Here are the facts.

BINOMIAL MEAN AND STANDARD DEVIATION

If a count X has the binomial distribution with number of observations n and probability of success p, the **mean** and **standard deviation** of X are

μ=npσ=np(1-p)

$$\begin{aligned} \mu &= np \\ \sigma &= \sqrt{np(1-p)} \end{aligned}$$

Beware that these short formulas are good only for binomial distributions. They can’t be used for other distributions.

EXAMPLE 12.6 Inspecting pharmaceutical containers

Continuing Example 12.5, the count X of containers contaminated with benzene is binomial with $n = 10$ and $p = 0.1$. The histogram in

Figure 12.2 displays this probability distribution. (Because probabilities are long-run proportions, using probabilities as the heights of the bars shows what the distribution of Xx would be in very many repetitions.) The distribution is strongly skewed. (The space between the bars indicates that counts can be only whole numbers). Although Xx can take any whole-number value from 0 to 10, the probabilities of values larger than 5 are so small that they do not appear in the histogram.

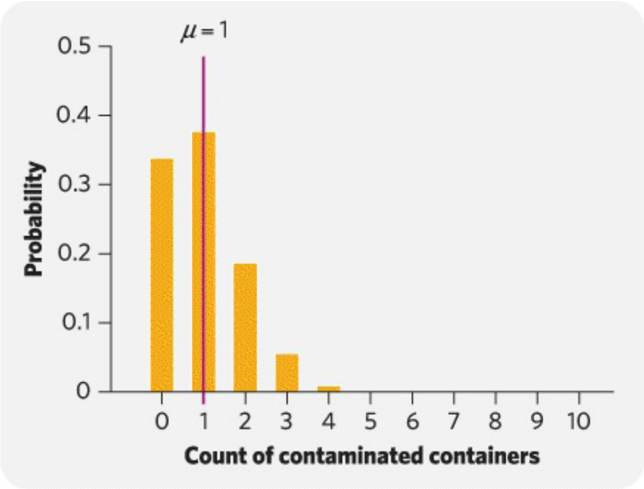


Figure 12.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 12.2 Probability histogram for the binomial distribution with n=10

$$n=10$$

and p=0.1.

$$p=0.1.$$

The mean and standard deviation of the binomial distribution in Figure 12.2 are

μ=np=(10)(0.1)=1σ=np(1-p)=(10)(0.1)(0.9)=0.9=0.9487

$$\begin{aligned}\mu &= np\\ &= (10)(0.1) = 1\\ \sigma &= \sqrt{np(1-p)}\\ &= \sqrt{(10)(0.1)(0.9)} = \sqrt{0.9} = 0.9487\end{aligned}$$

In many random samples of 10 containers, we would expect to find 1 container contaminated with benzene per sample, on average. The mean μμ is marked on the probability histogram in Figure 12.2. The standard deviation can be thought of, very roughly, as the expected deviation from this mean in many such random samples.

APPLY YOUR KNOWLEDGE

12.7 Antibiotic resistance, continued. Go back to Exercises 12.5 and 12.6 describing the percents of gonorrhea cases that were resistant to at least one of the three major antibiotics in 2010 and 2014, respectively.

a. In 2010, out of 10 randomly selected gonorrhea cases, what are the mean μμ and the standard deviation σσ of the count of cases XX showing resistance to at least one of the three major antibiotics?
b. Obtain the mean and standard deviation of XX in 2014.
c. How does the different value of the parameter pp in 2010 and 2014 affect the standard deviation σ?σ? What would be the value of σσ for n=10$n = 10$ if pp were equal to 0.5? What does your work suggest about the behavior of the standard deviation of a binomial distribution as the probability of a success gets closer to 0.5?

12.8 Pregnancy while on the birth control pill. The actual rate of unintended pregnancy for women taking the traditional birth control pill is 5%. That is, if a doctor prescribes the pill to 100 new female users, 5 of them will have become pregnant within their first year taking the pill. Therefore, the pill’s true efficacy under typical usage conditions is 95%.[5]

a. For 20 new female users, what is the mean number of unintended pregnancies? What is the mean number of women not experiencing any pregnancy within the first year of usage? Recognize that these two means must add to 20, the total number of women using the birth control pill.

b. What is the standard deviation σσ of the number XX of women not experiencing a pregnancy during the first year on the birth control pill? What is the standard deviation σσ of the number *Y* of women experiencing a pregnancy during that time? Notice that the standard deviation, which is a measure of variability, is the same regardless of which of the two outcomes we define as "success."

c. The birth control pill is often advertised as being 99% effective. In other words, under "perfect" laboratory conditions, the probability that a woman would not become pregnant over a usage period of one year is p=0.99.$p = 0.99.$ What is σσ in this case? What happens to the standard deviation of a binomial distribution as the probability of a "success" gets close to 1?

12.9 Mental health in schools. An *NPR* news article makes the following statement:[6] "Up to one in five kids living in the U.S. shows signs or symptoms of a mental health disorder in a given year. So in a school classroom of 25 students, five of them may be struggling with the same issues many adults deal with: depression, anxiety, substance abuse."

a. Explain why the example cited can be described in the context of the binomial setting.

b. The article cites three values (one in five, 25, five). What do they represent? Give both their mathematical symbol and their definition.

c. Explain why the second sentence cited could be misinterpreted. Rewrite this sentence so that it is both mathematically correct and understandable for the general public.

THE NORMAL APPROXIMATION TO BINOMIAL DISTRIBUTIONS

The formula for binomial probabilities becomes more challenging as the number nn of observations increases. You can use software or a graphing calculator to handle many problems for which applying the formula is not practical. In some situations, there is another alternative: *As the number of observations n gets larger, the binomial distribution gets close to a Normal distribution.* When nn is large, we can use Normal probability calculations to approximate binomial probabilities. We will see in Chapter 13 that this approximation has useful applications beyond just simplifying binomial computations.

EXAMPLE 12.7 Overweight Americans: binomial computation

According to the 2011 Health Interview Survey, 62% of American adults are either overweight or obese, based on their self-reported body mass index (BMI). Suppose that we take a random sample of 2000 adults. What is the probability that 1250 or more of the sample were overweight or obese in the year 2011?

Because there are more than 200 million adults in the United States, we can take the BMI values of 2000 randomly chosen adults to be independent observations. Then, the number in our sample who were either overweight or obese is a random variable XX having the binomial distribution with n=2000$n = 2000$ and p=0.62.$p = 0.62.$

To find the probability that at least 1250 of the people in the sample were overweight or obese, we must add the binomial probabilities of all outcomes from X=1250$X = 1250$ to X=2000.$X = 2000.$ This probability is P(X≥1250),$P(X \geq 1250),$ which the software R gives as 0.3313682 from the command 1 - pbinom (1249, size=2000, prob=0.62).

Performing the binomial calculations manually isn’t practical, but software gives an answer that is exactly correct to seven decimal places. Note that some programs give only cumulative probabilities, but you can make use of the fact that P(X≥1250)=1-P(X≤1249). $P(X \geq 1250) = 1 - P(X \leq 1249)$. Other programs, however, may not even perform the intensive calculations required for such a large n.n.

Let’s consider instead the Normal approximation to the binomial distribution. Figure 12.3 shows the actual binomial probability distribution for X,X, with events X≥1250$X \geq 1250$ in lighter color. The shape of this distribution closely resembles the Normal distribution with the same mean and standard deviation as the binomial variable X:X:

μ=np=(2000)(0.62)=1240σ=np(1-p)=(2000)(0.62)(0.38)=21.707

$$\begin{aligned} \mu &= np = (2000)(0.62) = 1240 \\ \sigma &= \sqrt{np(1-p)} = \sqrt{(2000)(0.62)(0.38)} = 21.707 \end{aligned}$$

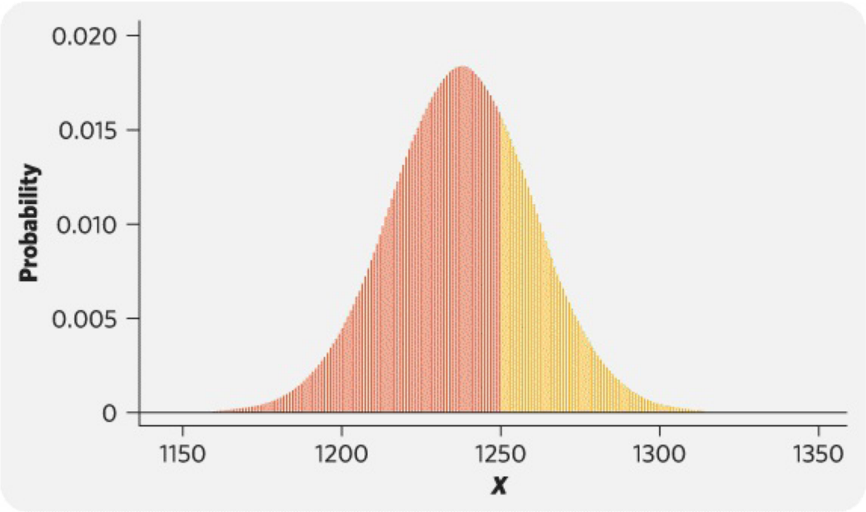
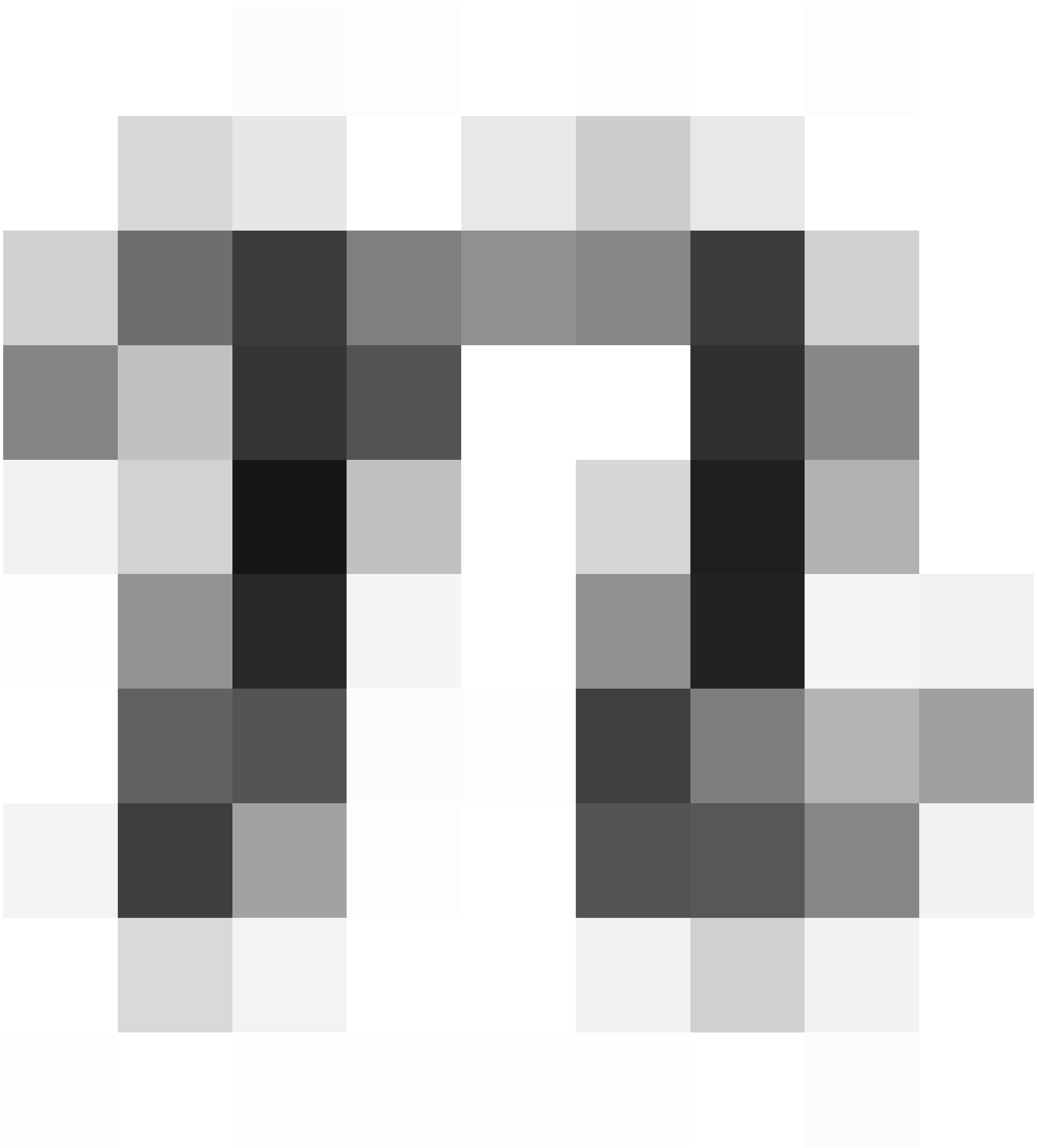


Figure 12.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

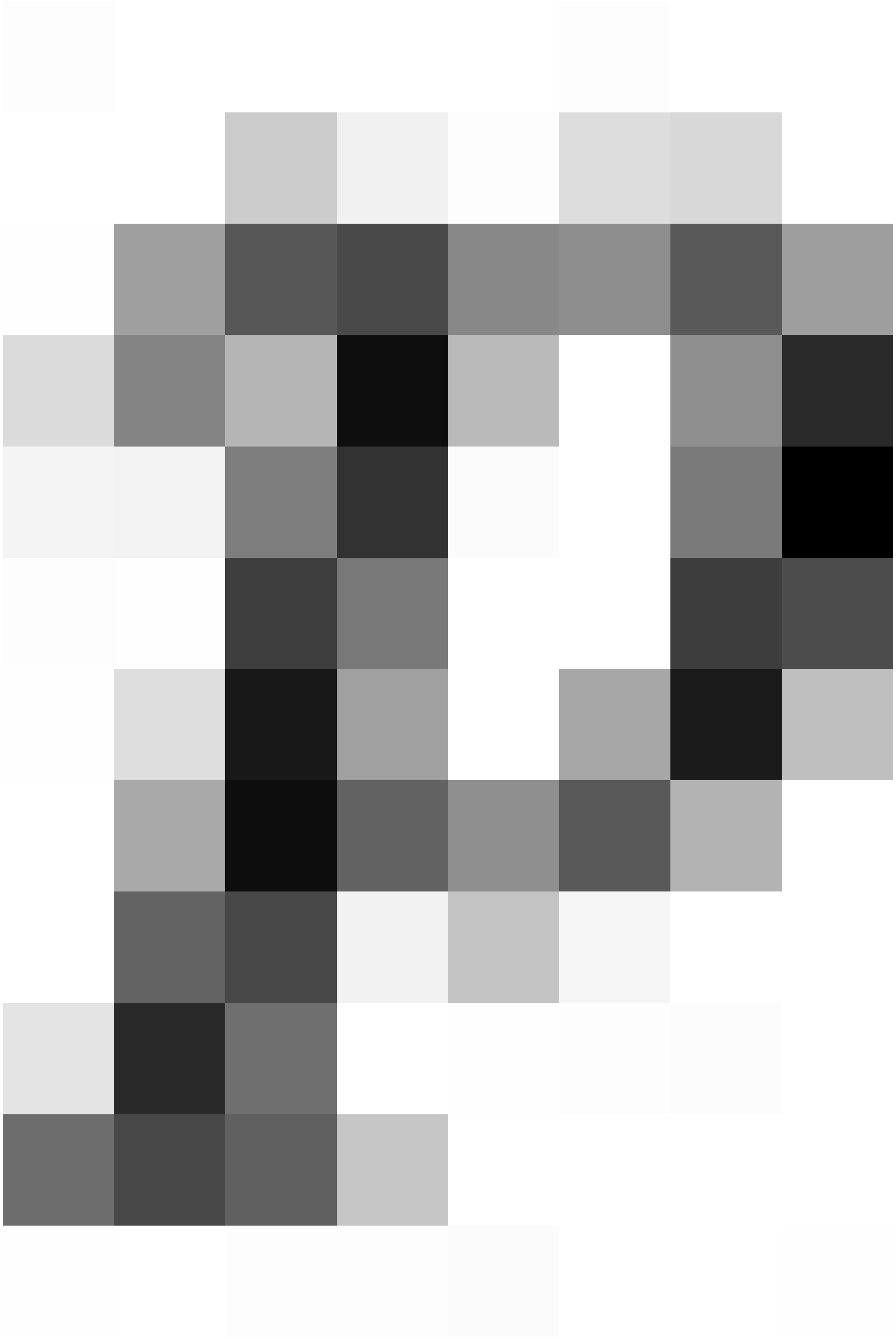
FIGURE 12.3 Probability distribution for the binomial model with n



=

[redacted]

2000 and p



=

0.62, displayed graphically. The height of each bar represents the probability for X

X

when it takes a value on the horizontal axis. The yellow bars represent probabilities for values of X≥1250.

$$X \geq 1250.$$

Notice how the shape of this binomial probability distribution closely resembles a Normal curve.

EXAMPLE 12.8 Overweight Americans: Normal approximation

When we act as though the count XX has the N(1240,21.707) $N(1240, 21.707)$ distribution, the software R gives P(X≥1250)=0.3225138 $P(X \geq 1250) = 0.3225138$ if we enter 1-pnorm(1250, mean = 1240, sd = 21.707). The Normal approximation differs from the binomial result 0.3313682 by only 0.009.

Here is P(X≥1250)$P(X \geq 1250)$ with the Normal approximation, using Table B instead:

P(X≥1250)=PX-124021.707≥1250-124021.707=P(Z≥0.46)=1-0.6772=0.3228

$$\begin{aligned} P(X \geq 1250) &= P\left(\frac{X-1240}{21.707} \geq \frac{1250-1240}{21.707}\right) \\ &= P(Z \geq 0.46) \\ &= 1 - 0.6772 = 0.3228 \end{aligned}$$

NORMAL APPROXIMATION FOR BINOMIAL DISTRIBUTIONS

Suppose that a count XX has the binomial distribution with nn observations and success probability p.p. When nn is large, the distribution of XX is approximately Normal, N(np,np(1-p)). $N(np, \sqrt{np(1-p)})$.

As a rule of thumb, we will use the Normal approximation when nn is so large that np≥10$np \geq 10$ and n(1-p)≥10. $n(1-p) \geq 10$.

The Normal approximation is easy to remember because it says that XX is Normal with its binomial mean and standard deviation. The accuracy of the Normal approximation improves as the sample size nn increases. It is most accurate for any fixed nn when pp is close to 0.5 and is least accurate when pp is near 0 or 1. This is why the rule of thumb in the box depends on pp as well as n.n. The *Normal Approximation to Binomial* applet shows in visual form how well the Normal approximation fits the binomial distribution for any nn and p.p. You can slide nn and watch the approximation get better. Whether you use the Normal approximation should depend on how

APPLET

accurate your calculations need to be. For most statistical purposes, great accuracy is not required. Our rule of thumb for use of the Normal approximation reflects this judgment.

For slightly more accurate results with the Normal approximation, you can use a **continuity correction**. Because counts can take only integer values but the Normal distribution can take any real value, the proper continuous equivalent to a count is the interval around it with size 1. For example, the continuous equivalent to a 1250 count is the interval between 1249.5 and 1250.5. Therefore, in Example 12.8 we would get a slightly more accurate Normal approximation of the probability that 1250 or more individuals in the sample were overweight or obese in the year 2011 by finding $P(X \geq 1249.5)$.$P(X \geq 1249.5)$. This gives probability 0.3308212 (obtained by entering 1-pnorm (1249.5, mean = 1240, sd = 21.707) in R), which is closer still to the binomial software output. The continuity correction is especially helpful when the sample size is small.

APPLY YOUR KNOWLEDGE

12.10 Checking for survey errors. One way of checking the effect of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known facts about the population. According to the 2015 U.S. census update, approximately 13% of Americans are black, or African American. The number XX of blacks in a random sample of 1500 Americans should therefore vary with the binomial (n=1500, p=0.13$n = 1500, p = 0.13$) distribution.

a. What are the mean and standard deviation of X?X?

b. Use the Normal approximation to find the probability that the sample will contain 200 or more blacks. Also check that you can safely use the approximation.

c. Use the 68–95–99.7 rule (on page 274) to find the approximate number of blacks that would correspond to the middle 95% of random samples of 1500 American adults. Verify your result using cumulative binomial probabilities.

12.11 Bone density. Forty percent of postmenopausal women have low bone density (osteopenia), placing them at risk for osteoporosis with ensuing spontaneous fractures. Osteoporosis is estimated to lead to $14 billion per year in medical expenses alone, yet it can be prevented if treated early enough. A physician group has 248 postmenopausal primary patients and has diagnosed 82 of these women with low bone density.

a. What is the probability that only 82 or fewer of a set of 248 patients have low bone density? Use the Normal approximation to find this probability.
b. Find the probability in part a using the binomial probability model. How does it compare with the Normal approximation?
c. Based on the probability of this outcome, what would you suggest at the next physicians meeting?

THE POISSON DISTRIBUTIONS

Binomial distributions are not the only type of discrete distribution for whole-number counts. Exercises 10.37, 12.51, and 12.52 show examples of geometric distributions, and Exercise 12.53 walks you through an example of a multinomial distribution. These distributions are beyond the scope of an introductory textbook and will not be discussed beyond these examples.

Poisson distributions are yet another class of discrete distributions and have important applications in the life sciences. They describe the count of events that occur at random over a chosen unit of space or time. For example, a Poisson distribution would be a good model for the count of chocolate chips in cookies made from the same well-mixed batch or for the number of times you walk up the stairs on any workday. Note that, unlike counts in a binomial distribution, counts in a Poisson distribution are open-ended because they have no fixed upper boundary: That is, they can, in theory, go on to infinity.

POISSON DISTRIBUTION

A **Poisson distribution** describes the count Xx of occurrences of a defined event in fixed, finite intervals of time or space when

1. Occurrences are all **independent** (that is, knowing that one event has occurred does not change the probability that another event may occur), and
2. The probability of an occurrence is the same over all possible intervals of the same size.

EXAMPLE 12.9 Counting dandelions

On a lazy summer afternoon, you lie in a field adorned with dandelions growing here and there. In an inquisitive spirit, you divide the field into quadrants of 1 square meter and count the number of dandelions in each quadrant. The number of dandelions per quadrant is the random variable X.X. If the dandelions present in each quadrant are random, independent events, then the distribution of XX follows approximately a Poisson distribution.

Figure 12.4 illustrates this concept via a simulation. The large rectangle represents a field divided into 50 quadrants of equal size. Each circle is a dandelions on the simulated field. The dandelions were randomly “seeded” according to a Poisson distribution, giving an overall mean of 2.1 dandelions per quadrant. Starting from the left, the first 10 quadrants on the upper row have, respectively, 2, 3, 2, 1, 3, 3, 0, 2, 2, and 2 dandelions each. Notice how the overall pattern of dandelions on the field appears to be entirely random.

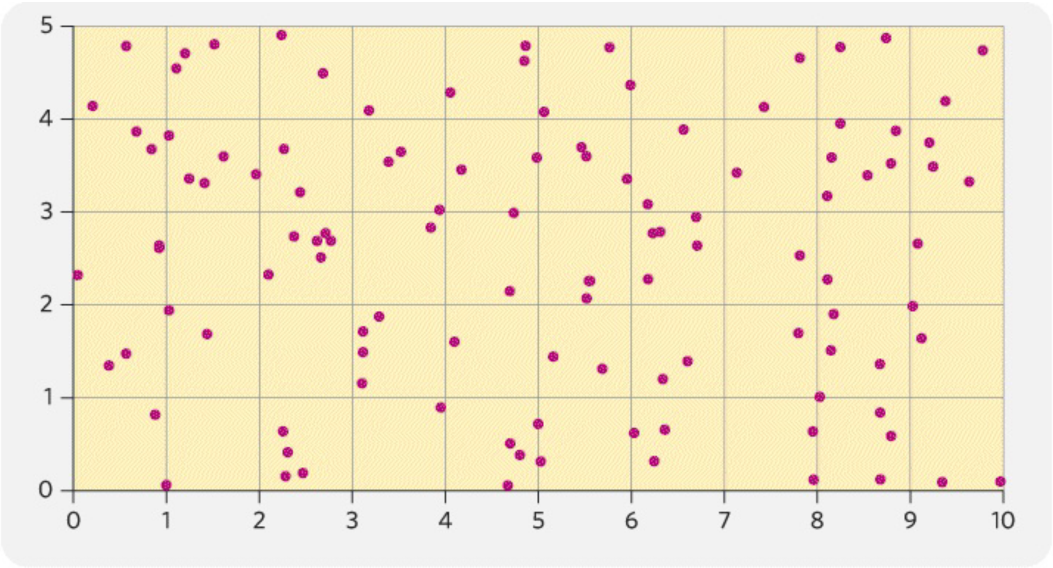


Figure 12.4

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

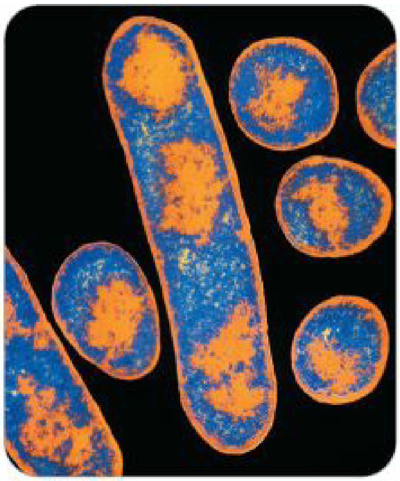
FIGURE 12.4 Graphical display of a simulated field containing dandelions (circles). Dandelion location on the rectangular field follows a Poisson distribution, with mean 2.1 dandelions per quadrant.

APPLY YOUR KNOWLEDGE

In Exercises 12.12 to 12.14, the random variable X is a count. Does X have a Poisson distribution? Give your reasons in each case.

12.12 Random digit dialing. When an opinion poll calls residential telephone numbers at random, only 20% of the calls reach a live person. You watch the random dialing machine make calls until a live person answers. Xx is the number of calls needed to reach a live person.

12.13 Botulism in California. Botulism is a rare but very serious bacterial infection caused by *Clostridium botulinum* acquired via contaminated food or infected wounds. It is one of several notifiable diseases that must be officially reported by doctors to state and federal health agencies. The California Department of Public Health reports the number Xx of botulism cases diagnosed in California each year.



A. Dowsett, Public Health England /Science Photo Library/Getty Images

12.14 Traffic accidents. The U.S. Department of Transportation tracks fatal vehicle crashes. We are interested in X,x, the number of fatal crashes that occur in the United States in any given one-hour interval.

POISSON PROBABILITIES

Poisson distributions can be used to predict what counts we would expect for a random event occurring within an interval of fixed size (in time or space). For instance, a medical group might model the count of non-life-threatening emergency room visits on weekend days and use the results to decide whether to open a walk-in clinic and how to staff it.

Because Poisson distributions describe random, independent events, it can also be interesting to find out if a random variable substantially deviates from that model. For instance, when the count of meningitis cases in a county exceeds what would reasonably be expected under a Poisson distribution, we can conclude that the meningitis cases are not randomly and independently distributed, suggesting instead that an epidemic outbreak is occurring.

To do this, we need a way to compute Poisson probabilities. Deriving mathematically the formula of a Poisson probability distribution is beyond the scope of an introductory statistics textbook.

POISSON PROBABILITY

If XX has the Poisson distribution with mean number of occurrences per interval μ,$\mu,$ the possible values of XX are 0, 1, 2, …. If *k* is any one of these values,

P(X=k)=e-μμkk!

$$P(X=k)=\frac{e^{-\mu}\mu^{k}}{k!}$$

The variance σ2σ^2 of the Poisson distribution is also equal to μ,$\mu,$ the mean number of occurrences per interval. The distribution’s standard deviation σσ is therefore equal to μ.$\sqrt{\mu}.$

Because the mean and the variance of a Poisson distribution are equal, when the mean number of occurrences is large, the variance is also large and the distribution looks very flat and wide. Therefore, Poisson distributions are most commonly used to describe rare, random events (or random events examined over small intervals).

EXAMPLE 12.10 Monitoring mumps outbreaks

Mumps is an acute viral infection that is generally mild and even asymptomatic in approximately 20% of infected individuals. However, complications can arise that have long-lasting consequences, such as meningitis/encephalitis, sterility, spontaneous abortion, or deafness. Mandatory vaccinations of infants have largely prevented mumps epidemics in the United States, with an average of only 265 mumps cases per year since 1996. In Iowa, the average monthly number of reported cases is approximately 0.1 over that same period.[7]

Dr P. Marazzi/Science Source

If we assume that cases of mumps are random and independent, the number XX of monthly mumps cases in Iowa has approximately the Poisson distribution with μ=0.1.$\mu = 0.1$.

The probability that in a given month there is no more than 1 mumps case is

P(X≤1)=P(X=0)+P(X=1)=e-0.1×0.100!+e-0.1×0.111!=0.9048×11+0.9048×0.11=0.9048+0.0905=0.9953

$$\begin{aligned} P(X \le 1) &= P(X=0) + P(X=1) \\ &= \frac{e^{-0.1} \times 0.1^0}{0!} + \frac{e^{-0.1} \times 0.1^1}{1!} \\ &= \frac{0.9048 \times 1}{1} + \frac{0.9048 \times 0.1}{1} \\ &= 0.9048 + 0.0905 = 0.9953 \end{aligned}$$

We would expect no more than 1 case of mumps per month with probability 99.53%. That is, we would consider seeing two or more mumps cases in a month a very unlikely event (probability 1-0.9953 $1-0.9953$) if cases were random and independent with a constant monthly

> rate of 0.1.

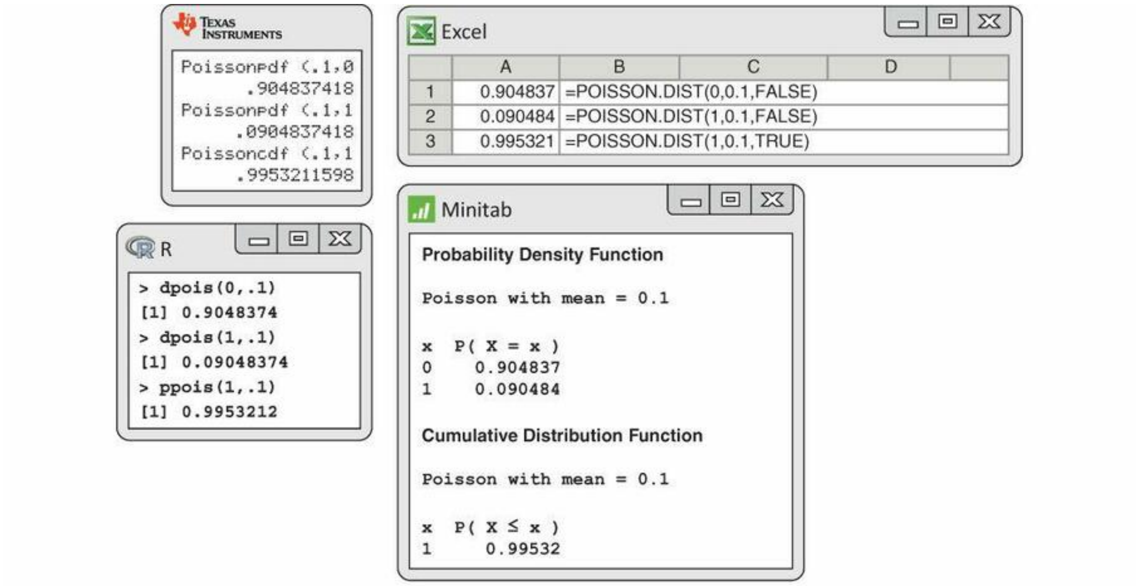
STATS IN YOUR WORLD

MUMPS IN PERSPECTIVE

According to the Centers for Disease Control and Prevention, before approval of the mumps vaccine in 1967, nearly everyone in the United States experienced mumps, with about 200,000 new cases reported every year. The mumps vaccine has a reported effectiveness of about 95%. That is, some individuals will not be helped by the vaccine because they lack the genetic ability to produce antibodies against that particular viral strain.

The Poisson probability formula is simple to use for any value of X,X, but it can be more cumbersome for cumulative probabilities. Again, technology is a better choice.

Figure 12.5 shows the output for the calculation in Example 12.10 from the TI-83 graphing calculator and three computer programs. All four outputs agree with the result 0.9953, up to roundoff error, whether it is obtained by adding P(X=1)$P(X=1)$ and P(X=0)$P(X=0)$ or by directly computing the cumulative probability P(X≤1).$P(X \leq 1)$. If all our assumptions are correct, the probability that Iowa would see one or fewer mumps cases in any given month is about 0.995.



```
Poissonpdf (.1,0
         .904837418
Poissonpdf (.1,1
        .0904837418
Poissoncdf (.1,1
        .9953211598
```

	A	B	C	D
1	0.904837	=POISSON.DIST(0,0.1,FALSE)		
2	0.090484	=POISSON.DIST(1,0.1,FALSE)		
3	0.995321	=POISSON.DIST(1,0.1,TRUE)		

```
> dpois(0,.1)
[1] 0.9048374
> dpois(1,.1)
[1] 0.09048374
> ppois(1,.1)
[1] 0.9953212
```

```
Probability Density Function

Poisson with mean = 0.1

x  P( X = x )
0     0.904837
1     0.090484

Cumulative Distribution Function

Poisson with mean = 0.1

x  P( X ≤ x )
1      0.99532
```

Figure 12.5

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 12.5 The Poisson probability P(X≤1):

$$P(X \le 1):$$

output from three computer programs and a graphing calculator.

While the exact function names and layout vary across platforms, all four technologies in Figure 12.5 require the following information: the mean μ,μ, the value of x,x, and an indication of whether the probability P(X=x)$P(X = x)$ or the cumulative probability P(X≤x)$P(X \le x)$ is desired. Comparing Figures 12.1 and 12.5, we notice that each specific technology platform is consistent in how it differentiates a simple probability for one value of xx and a cumulative probability (for example, “pdf” versus “cdf” for the TI-83 calculator).

EXAMPLE 12.11 Monitoring mumps outbreaks

Mumps cases, like all cases of rare and communicable diseases, must be reported to the CDC and data on their occurrence are published in *Morbidity and Mortality Weekly Report*. In January 2006, Iowa reported 4 cases of mumps. What is the probability of getting 4 or more cases of mumps under the model used in Example 12.10?

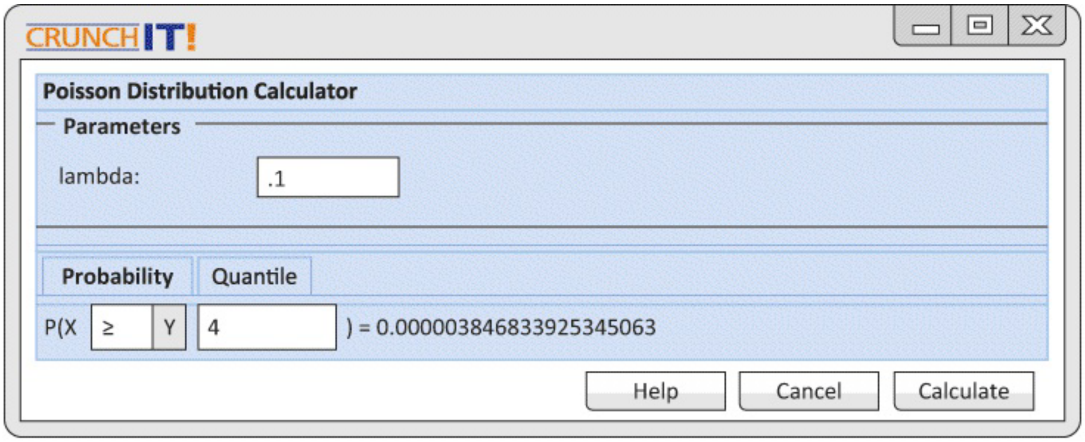
The probability that in a given month there are 4 or more mumps cases is

$$P(X \geq 4)=1-P(X \leq 3)=1-0.999996=0.000004$$

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - 0.999996 = 0.000004 \end{aligned}$$

Using the TI-83 calculator, we find that the function poissoncdf(0.1,3) for $P(X \leq 3)$$P(X \leq 3)$ returns the value 0.9999961532. Some statistical programs are more flexible than others, letting you pick directly the probability you need. Figure 12.6 shows the output from the statistical software CrunchIt!, which gives $P(X \geq 4)=0.0000038\ldots.$

$P(X \geq 4) = 0.0000038\ldots.$



lambda: .1

P(X ≥ 4) = 0.000003846833925345063

Figure 12.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 12.6 CrunchIt! output for the Poisson probability $P(X \geq 4)$.

$$P(X \geq 4).$$

Assuming that cases of mumps are random and independent, we would expect to almost never see 4 or more cases of mumps in Iowa in any given month. The unusually high count of mumps cases points to a substantial departure from the Poisson model—for instance, a contagious outbreak. In a contagion model, knowing that one person is infected

> increases the chance that another person also is infected. Therefore, contagion events are not independent.

These first 4 mumps cases were indeed the beginning of a major mumps outbreak in Iowa, with the number of new cases increasing every week, totaling 219 cases by the end of March and 2597 cases by the end of April 2006, spread over several states. The mumps strain in the Iowa epidemic was genotype G, the same strain that has been responsible for other epidemic outbreaks in the United States and the United Kingdom, even in vaccinated individuals.[8] The CDC tracks all reports of rare diseases to understand their origin and prevent their spread and future recurrence.

Biological data should not be expected to exactly *match a mathematical model. This is a common misconception. Rather mathematical models are reasonable representations for various biological features.* Plant and animal distributions, for instance, can also be compared with a Poisson distribution to determine whether the events are truly random and independent. In Example 12.9 you looked at the distribution of dandelions in field quadrants. Plants disseminated by animals or by wind (such as dandelions) are expected to have geographical distributions that can be approximated by a Poisson model. In contrast, plants disseminated primarily asexually—for example, through underground rhizomes capable of sprouting new clones (such as lily of the valley)—are expected to be more clustered than would be predicted by a Poisson model.

APPLY YOUR KNOWLEDGE

12.15 Down syndrome. The state of New York reported 1484 live births in which the infants had Down syndrome (trisomy 21) between 2006 and 2010, which averages to about 5.7 cases per week.[9] While the causes of Down syndrome are not fully understood, it is reasonable at this point to assume that live births are independent and the weekly rate is constant. Let Xx be the count of babies born with Down syndrome in the state of New York in a given week.

a. What values could Xx take? Notice how there is no clear limit to this range.
b. What distribution does Xx follow, approximately?
c. Give the mean and standard deviation of X.x.

DenysKuvaiev/Depositphotos.com

12.16 Typhoid fever. The CDC receives reports of 7.7 cases of typhoid fever per week, on average, from all over the United States, although most cases were acquired during international travel. Typhoid fever is a life-threatening bacterial illness caused by *Salmonella typhi* and is transmitted by ingestion of contaminated food or drinks.[10]

a. What distribution would best represent the distribution of weekly counts, X,x, of typhoid fever?
b. What are the mean and standard deviation of this distribution?

12.17 Down syndrome, continued. In the setting of Exercise 12.15, in the state of New York:

a. What is the probability that no child is born with Down syndrome in a given week?
b. What is the probability that at least 1 child is born with Down syndrome in a given week?
c. What is the probability that 10 or more children are born with Down syndrome in a given week?

12.18 Typhoid fever, continued. In the setting of Exercise 12.16:

a. Use technology to find the cumulative probabilities for Xx equals

0, 1, 2, 3, to 15.

b. What is the probability that the CDC receives reports of more than 15 cases of typhoid fever next month? If this were to happen, what would you conclude?

CHAPTER 12 SUMMARY

- Acount XX of successes has a **binomial distribution** in the **binomial setting:** There are nn observations, the observations are independent of each other, each observation results in a success or a failure, and each observation has the same probability pp of a success.
- The binomial distribution with nn observations and probability pp of success gives a good approximation of the sampling distribution of the count of successes in an SRS of size nn from a large population containing proportion pp of successes.
- If XX has the binomial distribution with parameters nn and p,$p,$ the possible values of XX are the whole numbers 0, 1, 2, …, n.$n.$ The **binomial probability** that XX takes any value is

  P(X=k)=nkpk(1-p)n-k

  $$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

  In practice, binomial probabilities and cumulative probabilities are best found using software.
- The **binomial coefficient**

  nk=n!k!(n-k)!

  $$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

  counts the number of ways *k* successes can be arranged among nn observations. Here the **factorial** n!$n!$ is

  n!=n×(n-1)×(n-2)×···×3×2×1

  $$n!=n\times(n-1)\times(n-2)\times\cdots\times3\times2\times1$$

  for positive whole numbers n,$n,$ and 0!=1.$0!=1.$
- The **mean** and **standard deviation** of a binomial count XX are

  μ=npσ=np(1-p)

  $$\begin{aligned}\mu &= np\\ \sigma &= \sqrt{np(1-p)}\end{aligned}$$

- The **Normal approximation** to the binomial distribution says that if XX is acount having the binomial distribution with parameters nn and p,$p,$ then when nn is large, the distribution of XX is approximately N(np,np(1-p)).$N(np, \sqrt{np(1-p)})$. It is best to use this approximation only when np≥10$np \geq 10$ and n(1-p)≥10.$n(1-p) \geq 10.$

- The **Poisson distribution** with mean μμ describes the count XX of occurrences of random, independent events in fixed intervals of time or space when the probability of an event occurring is the same over all possible such intervals.

- If XX has the Poisson distribution with mean number of occurrences per interval μ,$\mu,$ the possible values of XX are 0, 1, 2, …. The **Poisson probability** that XX takes any value is

  P(X=k)=e-μμkk!

  $$P(X=k)=\frac{e^{-\mu}\mu^{k}}{k!}$$

  In practice, Poisson probabilities and cumulative probabilities are best found using software.

- The **mean** and **standard deviation** of a Poisson count XX are

  μ=mean number of occurrences per intervalσ=μ

  $$\begin{aligned}\mu &= \text{mean number of occurrences per interval}\\ \sigma &= \sqrt{\mu}\end{aligned}$$

THIS CHAPTER IN CONTEXT

The binomial distribution is used to compute probabilities for the count of successes among n observations that are produced under the binomial setting. An important situation in which the binomial setting can be used is when we choose an SRS from a large population with a proportion p of successes, as in the context of sample surveys discussed in Chapter 6. Calculations can be performed by using the binomial formula. Alternatively, when n and p are such that both the mean number of successes and the mean number of failures are large enough, the Normal approximation can be used.

Another important application of the binomial distribution is in making inferences about the unknown *proportion p* of some outcome in a population. The proportion of some outcome in an SRS is simply the count of such outcomes divided by the total number of observations n. The methodology used for inference about a proportion is a direct consequence of the Normal approximation for the binomial distribution, as we will see in Chapter 13 when discussing sampling distributions. In Chapter 19 we will examine how to use the sampling distribution of a proportion to estimate the unknown population parameter (the true proportion of an outcome of interest in a given population) or to test a hypothesis about it.

The Poisson distribution is used to compute probabilities for the count of rare random events in fixed intervals of time or space when the events are independent and occur with constant mean rate μ. Poisson probabilities are a good model for completely random events. In Chapter 21 we will see an example of statistical inference using the chi-square test for goodness of fit to test the hypothesis that a random event has a Poisson distribution.

CHECK YOUR SKILLS

Tay-Sachs disease is a fatal, recessive genetic disorder in which harmful quantities of a fatty waste substance accumulate in the brain, with the condition slowly progressing until it leads to death by age five. When both parents are carriers of the Tay-Sachs mutation, the probability is 0.25 that their child will have Tay-Sachs disease. Successive children inherit genetic material from their parents independently. Use this information to answer Exercises 12.19 and 12.20.

12.19 For a couple in which both parents carry the Tay-Sachs mutation, the number of children with Tay-Sachs disease that they could have if they want 3 children (ignore the possibility of identical twins) has the distribution

a. binomial with n=4$n = 4$ and p=1/4.$p = 1/4.$

b. binomial with n=3$n = 3$ and p=1/4.$p = 1/4.$

c. binomial with n=3$n = 3$ and p=1/3.$p = 1/3.$

12.20 What is the probability that at least 1 of their 3 children will have Tay-Sachs disease?

a. 0.68

b. 0.58

c. 0.30

12.21 In a group of 10 college students, 6 have never smoked tobacco regularly. You choose 3 of the 10 students at random and ask about their smoking status. The distribution of the number you choose who have never smoked regularly is

a. binomial with n=10$n = 10$ and p=0.6.$p = 0.6.$

b. binomial with n=3$n = 3$ and p=0.6.$p = 0.6.$

c. not binomial.

The U.S. National Center for Health Statistics reports that approximately 12% of emergency department visits result in hospital admissions.[11] *Consider 20 randomly selected emergency department visits and assume* *that visits to emergency departments are independent. Exercises 12.22 to 12.25 use this setting and the following Minitab output:*

Binomial with n=20$n = 20$ and p=0.12
$p = 0.12$

xx	P(X=x)$P(X = x)$	P(X≤x)$P(X \le x)$
0	0.077563	0.077563
1	0.211535	0.289098
2	0.274034	0.563132
3	0.224210	0.787341

12.22 How many hospital admissions do we expect, on average, in random samples of 20 emergency department visits?

a. 2
b. 2.4
c. 12

12.23 What is the approximate probability that exactly 2 of the 20 visits would result in hospital admissions?

a. 0.0144
b. 0.2740
c. 0.5631

12.24 What is the approximate probability that at most 2 of the 20 visits would result in hospital admissions?

a. 0.1344
b. 0.2740
c. 0.5631

12.25 If XX is the number of emergency department visits that result in hospital admissions in random samples of 20 visits, what is approximately the standard deviation σσ of X?$X?$

a. 1.45
b. 2.11
c. 4.46

12.26 The mean number of daily surgeries at a local hospital is 6.2. If we assume that surgeries are random, independent events, the count of daily surgeries follows approximately

a. a binomial distribution, with mean 6.2 and standard deviation 3.8.
b. a Poisson distribution, with mean 6.2 and standard deviation 6.2.
c. a Poisson distribution, with mean 6.2 and standard deviation 2.49.

12.27 The mean number of daily surgeries at a local hospital is 6.2. If we assume that surgeries are random, independent events, what is the approximate probability that there would be only 2 or fewer surgeries in a given day?

a. 0.015
b. 0.039
c. 0.054

12.28 Which of the following would you expect to follow approximately a Poisson distribution?

a. The weekly count of toddlers with a runny nose at a day care
b. The count of adults with a disability visiting the post office daily
c. The count of children with color blindness in a classroom of 20

CHAPTER 12 EXERCISES

12.29 Binomial setting? In each of the following situations, is it reasonable to use a binomial distribution for the random variable X?x? Give reasons for your answer in each case.

a. A quality controller for medical catheters randomly chooses one catheter each hour during an 8-hour production shift for a detailed quality inspection. One variable recorded is the count Xx of catheters with an unacceptable diameter (too small or too large).
b. A sample survey asks 100 persons chosen at random from the adult residents of a large city whether they oppose the legalization of medical marijuana; Xx is the number who say “yes.”
c. A pediatrician sees 24 unrelated children on one winter day; Xx is the number of patients who came because of cold or flu symptoms.

12.30 Binomial setting? Your pulse reflects the rate at which your heart beats. A healthy resting pulse rate in adults is between 60 and 100 beats per minute. In which of these two settings is a binomial distribution more likely to be at least approximately correct? Explain your answer.

a. You select at random 10 times in the next 24-hour period to take your pulse. You then go about your usual activities, including exercising and sleep. The random variable Xx is the number of measurements that are less than 100 beats per minute.
b. A nurse instructs an adult patient to take his resting pulse 10 times over the next few days. A resting pulse requires resting for at least 10 minutes before taking the pulse. The random variable Xx is the number of measurements that are less than 100 beats per minute.

12.31 Testing ESP. In a test for ESP (extrasensory perception), a subject is told that cards only the experimenter can see contain either a star, a circle, a wave, or a square. As the experimenter looks at each of 20 cards in turn, the subject names the shape on the card. A subject who is just guessing has probability 0.25 of guessing correctly on each card.

a. The count of correct guesses in 20 cards has a binomial distribution. What are n and p?

b. What is the mean number of correct guesses in 20 cards for subjects who are just guessing?

c. What is the probability of exactly 5 correct guesses in 20 cards if a subject is just guessing?

12.32 Genetics. According to genetic theory, the blossom color in the second generation of a certain cross of sweet peas should be red or white in a 3:1 ratio. That is, each plant has probability 3/4 of having red blossoms, and the blossom colors of separate plants are independent.

a. What is the probability that exactly 6 out of 8 of these plants have red blossoms?

b. What is the mean number of red-blossomed plants when 80 plants of this type are grown from seeds?

c. What is the probability of obtaining at least 50 red-blossomed plants when 80 plants are grown from seeds?

12.33 Unvaccinated school-age children. A study of immunizations among school-age children in California found that some areas have rates of unvaccinated school-age children as high as 13.5%.[12] Let X be the number of children not vaccinated in a randomly selected classroom of 30 children in one such area.

a. What is the distribution of X?

b. How many unvaccinated children do we expect, on average, in randomly selected classrooms of 30 children?

c. What is the probability that there is at least 1 unvaccinated child in a classroom of 30?

d. What is the probability that a classroom of 30 children has 5 or more unvaccinated children?

12.34 Polydactyly. In the United States, 1 in every 500 babies is born with polydactyly, a harmless congenital abnormality in which a newborn has one or more extra fingers or toes. Cases occur at random and independently, without apparent cause. Let X be the count of babies born with polydactyly in a random sample of 1000 births.

a. What distribution does XX follow? Give the mean and standard deviation of X.$X.$
b. What is the probability that no child was born with polydactyly in such a sample? Exactly 2 children?
c. What is the probability that 5 or more children were born with polydactyly in such a sample?

12.35 Malaria epidemics. Malaria is a debilitating and potentially deadly parasitic infection that is mostly preventable and curable, yet many developing countries lack the necessary resources to manage it. The West African country of Guinea has the highest rate of malaria in the world, with 75% of its population being infected.[13] A physician from Doctors Without Borders sees 6 unrelated patients in a given hour. Let XX be the count of patients who are infected with malaria among the 6 patients the doctor saw.

a. The count XX has a binomial distribution. What are nn and p?$p?$
b. What are the possible values that XX can take?
c. Find the probability of each value of X.$X.$ Draw a probability histogram for the distribution of X.$X.$ (See Figure 12.2 for an example of a probability histogram.)
d. What are the mean and standard deviation of this distribution? Mark the location of the mean on your histogram.

12.36 False-positives in testing for HIV. A rapid test for the presence in the blood of antibodies to human immunodeficiency virus (HIV), the virus that causes acquired immunodeficiency syndrome (AIDS), gives a positive result with probability about 0.004 when a person who is free of HIV antibodies is tested. Consider rapid tests performed on 1000 people who are all free of HIV antibodies.

a. What is the distribution of the number of positive tests?
b. What is the mean number of positive tests?
c. You cannot safely use the Normal approximation for this distribution. Explain why.

12.37 Inheriting blood type. If two parents each had exactly one parent with blood type O, then there is probability 0.25 that any of their children would have blood type O. Consider such parents who together have 5 children. The number of such children who have type O blood is a random variable XX that has the binomial distribution with n=5$n = 5$ and p=0.25. $p = 0.25.$

a. What are the possible values of X?$X?$
b. Find the probability of each value of X.$X.$ Draw a histogram to display this distribution like the one in Figure 12.2. (Because probabilities are long-run proportions, a histogram with the probabilities as the heights of the bars shows what the distribution of XX would be in very many repetitions.)
c. What are the mean and the standard deviation of X?$X?$ Mark the location of the mean on the probability histogram you made in that exercise.

Whooping cough (pertussis) is a highly contagious bacterial infection that was a major cause of childhood deaths before the development of vaccines. Approximately 80% of unvaccinated children who are exposed to whooping cough will develop the infection, compared to only 5% of vaccinated children. Exercises 12.38 to 12.41 are based on this information.

12.38 Vaccination at work. A group of 20 children at a nursery school are exposed to whooping cough by playing with an infected child.

a. If all 20 children have been vaccinated, what is the mean number of new infections? What is the probability that no more than 2 of the 20 children will develop infections?
b. If none of the 20 has been vaccinated, what is the mean number of new infections? What is the probability that 18 or more of the 20 children will develop infections?

12.39 A whooping cough outbreak. In 2007, Bob Jones University ended its fall semester a week early because of a whooping cough outbreak; 158 students were isolated with suspected infection and another 1200 given antibiotics as a precaution.[14] Authorities react strongly to whooping cough outbreaks because the disease is so contagious. Because the effect of childhood vaccination often wears off by late adolescence, treat the Bob Jones students as if they were unvaccinated. It appears that approximately 1400 students were exposed. What is the probability that at least 75% of these students would develop infections if not treated? (Fortunately, whooping cough is much less serious after infancy.)

12.40 A mixed group: means. A group of 20 children at a nursery school are exposed to whooping cough by playing with an infected child. Of these children, 17 have been vaccinated and 3 have not.

a. What is the distribution of the number of new infections among the 17 vaccinated children? What is the mean number of new infections?
b. What is the distribution of the number of new infections among the 3 unvaccinated children? What is the mean number of new infections?
c. Add your means from parts a and b. This is the mean number of new infections among all 20 exposed children.

12.41 A mixed group: probabilities. We would like to find the probability that exactly 2 of the 20 exposed children in Exercise 12.40 will develop whooping cough.

a. One way to get 2 children with infections is to get 1 case among the 17 vaccinated children and 1 case among the 3 unvaccinated children. Find the probability of exactly 1 infection among the 17 vaccinated children. Find the probability of exactly 1 infection among the 3 unvaccinated children. These events are independent: What is the probability of exactly 1 infection in each group?
b. Write down all the ways in which 2 children with infections can be divided between the two groups of children. Follow the pattern of part a to find the probability of each of these possibilities. Add all your results, including the result of part a, to obtain the probability of exactly 2 children with infections among the 20 children.

12.42 The continuity correction. One reason why the Normal approximation may fail to give accurate estimates of binomial probabilities is that binomial distributions are discrete and Normal distributions are continuous. That is, counts take only whole-number values, but Normal variables can take any value. We can improve the Normal approximation by treating each whole-number count as if it occupied the interval from 0.5 below the number to 0.5 above the number. For example, you can approximate a binomial probability P(X≥10)$P(X \geq 10)$ by finding the Normal probability P(X≥9.5).$P(X \geq 9.5)$. Be careful: Binomial P(X>10) $P(X > 10)$ is approximated by Normal P(X≥10.5).$P(X \geq 10.5)$.

In Exercise 12.3 we learned that 44% of high school students in New York have ever had sexual intercourse. The exact binomial probability that 50 or more students in an SRS of 100 high school students from New York have ever had sexual intercourse is 0.1341.

a. Show that this setting satisfies the rule of thumb for use of the Normal approximation.
b. What is the Normal approximation to P(X≥50)?$P(X \geq 50)$?
c. What is the Normal approximation using the continuity correction? This value is much closer to the true binomial probability.

12.43 Monitoring rabies in Florida. Rabies is an often-fatal disease, which is typically transmitted through the bite of an infected animal. The state of Florida has been recording all known cases of rabies for the past 20 years, with data indicating an average rate of 3.6 cases of rabies per week.[15] Let XX be the count of rabies cases reported in a given week.

a. What distribution does XX follow, approximately? Give the mean and standard deviation of X.X.
b. What is the probability that there would be no case of rabies reported in a given week? What is the probability that there would be at least 1 case reported in a given week?

12.44 Blood plasma donations. Blood type AB is somewhat rare in the United States (it is found in 4% of the general population). It is particularly useful for plasma transfusions because this plasma can be accepted by individuals with all blood types. The plasma is the liquid part of the blood once all cells are removed. You count the number XX of AB donors who come to a fixed Red Cross location per day. This count has the Poisson distribution, with parameter μμ being the mean number of AB donors per day.

a. This donation center sees an average of 2.1 AB donors per day. Find the Poisson probabilities for X=0,1,2,3,4,5.$X = 0, 1, 2, 3, 4, 5$.

b. What is the probability that this particular location will get more than 5 AB donors in one day?

12.45 Planning for hospital admissions. A private hospital gets an average of 3.2 hospital admissions per day from its emergency room. XX is the count of hospital admissions from the emergency room in a day. To plan for the number of beds that should be kept available, you compute the probabilities P(X≥5),$P(X \geq 5)$, P(X≥6),$P(X \geq 6)$, and P(X≥7).$P(X \geq 7)$. What do you conclude?

12.46 Salmonellosis outbreak. Improperly handled or undercooked poultry and eggs are the most frequent cause of *Salmonella* food poisoning. Salmonellosis can be deadly, although most individuals recover on their own, without treatment. Cases must be reported to the CDC, where they are closely monitored. The average number of salmonellosis cases per month in South Dakota is 1.67.

a. If XX is the monthly count of salmonellosis cases in South Dakota, what distribution does XX follow, approximately? Give the mean and standard deviation of X.X.

b. Give the probabilities P(X=0),$P(X = 0)$, P(X≤1),$P(X \leq 1)$, P(X≤2),$P(X \leq 2)$, P(X≤3),$P(X \leq 3)$, and P(X≤4).$P(X \leq 4)$. What is the probability that there would be more than 4 cases of salmonellosis in South Dakota in a given month?

c. In September 1994, 14 cases of salmonellosis in South Dakota were reported to the CDC.[16] What is the probability that 14 or more cases would arise in one month in South Dakota? What does this unusual report suggest? If you worked at the CDC at that time, what do you think would be your next action?

12.47 Salmonellosis outbreak, continued. Refer to Exercise 12.46. The CDC reports that the average number of salmonellosis cases per month in Wisconsin is 15.58.

a. If XX is the monthly count of salmonellosis cases in Wisconsin, what distribution does XX follow, approximately? Give the mean and standard deviation of X.X.

b. Use technology to find the probabilities P(X=0),$P(X = 0)$, P(X≤5), $P(X \leq 5)$, P(X≤15),$P(X \leq 15)$, and P(X≤25).$P(X \leq 25)$. What is the probability that there would be more than 25 cases of salmonellosis in Wisconsin in a given month?

c. As in South Dakota, there were an unusually high number of salmonellosis cases in Wisconsin in September 1994, with 48 cases occurring that month. What is the probability that 48 or more cases would arise in one month in Wisconsin? What do you make of the fact that a salmonellosis outbreak occurred in two states at the same time? In fact, the salmonellosis outbreak was traced back to nationally distributed ice cream products from one factory. The factory was shut down until cleared, and its contaminated products recalled, thereby preventing many more salmonellosis cases.

12.48 A new mumps outbreak. For the week of February 6, 2010, the CDC reported a total of 68 cases of mumps, whereas the historical weekly average was only 11 cases, based on data from the past five years.[17]

a. Assuming that cases of mumps are random and independent, the number XX of weekly mumps cases in the United States has approximately what distribution? Given this assumption, what are the mean and standard deviation of X?X?

b. Based on this model, use technology to find the probabilities P(X≤15), $P(X \leq 15)$, P(X≤30),$P(X \leq 30)$, and P(X≤67).$P(X \leq 67)$. What is the probability that there would be 68 or more cases of mumps in one week in the United States?

c. Do these data from February 2010 support the idea that mumps cases are random and independent, with a mean of 11 cases per week? Explain your answer. In fact, these cases were part of the largest mumps outbreak in the United States since the 2006 outbreak described in Example 12.11, with more than 1500 cases of mumps diagnosed as of February 28, 2010. The outbreak began in June 2009 with the return of an 11-year-old boy from a trip to the United Kingdom while that country was experiencing a local mumps epidemic.

12.49 Earthquakes in Oklahoma. Oklahoma is a Great Plains state in the south-central part of the United States, a region not historically known for its earthquakes. Seismic records up to 2008 show that Oklahoma has experienced a constant rate about 1.5 perceptible earthquakes (magnitude at least 3) per year, on average.[18]

a. Assuming that earthquakes are random and independent, with a constant rate of 1.5 per year, the count of perceptible earthquakes per year in Oklahoma should have a Poisson distribution with mean 1.5. Under this model, what is the probability that Oklahoma had more than 10 perceptible earthquakes in a given year prior to 2008?
b. In 2013, Oklahoma experienced 109 perceptible earthquakes (an average of about 2 events per week). If the rate of 1.5 perceptible earthquakes per year was still true, what would be the probability of recording 109 or more perceptible earthquakes?
c. In 2014, Oklahoma experienced 585 perceptible earthquakes.[19] If the rate of 1.5 perceptible earthquakes per year was still true, what would be the probability of recording 585 perceptible earthquakes or more?
d. Based on the probabilities you computed in parts a through c, would you conclude that the mean number of perceptible earthquakes per year in Oklahoma has increased? Explain your answer.

12.50 Earthquakes in Oklahoma, continued. Natural oil and gas extraction brings up large quantities of contaminated water, most of which is then stored in deep disposal wells. The newer hydraulic fracturing process (fracking) uses high-pressure fluid injections to release oil and gas, creating additional water waste. Scientific studies indicate that fracking activity, and especially the lubricating effect of disposal wells on existing faults, can trigger earthquakes.[20] Oklahoma is one state that has recently seen a huge development of its fracking operations.

In 2015, Oklahoma experienced 907 perceptible earthquakes (far surpassing California[21]), for an average of about 2.5 perceptible earthquakes per day. Consider perceptible earthquakes that occur independently with a constant daily rate of 2.5.

a. What is the probability of a day with no perceptible earthquakes?

b. What is the probability that a given day would have 1 perceptible earthquake?

c. What is the probability that a given day would have 5 or more perceptible earthquakes?

Exercises 12.51 through 12.53 explore other well-known discrete probability distributions not described in the core of this chapter. They are optional material.

12.51 Osteoporosis. Osteoporosis is a disease defined as very low bone density resulting in a high risk of fracture, hospitalization, and immobilization. Seven percent of postmenopausal women have osteoporosis.[22] A physician administers routine check-ups. Visits by postmenopausal women are independent. We are interested in how long the physician must wait before encountering the first postmenopausal woman with osteoporosis.

a. Considering only patient visits by postmenopausal women, the probability of a patient with osteoporosis on the first visit is 7%. What is the probability that the first patient does not have osteoporosis and the second patient does?

b. What is the probability that the first two patients do not have osteoporosis and the third one does? This is the probability that the first patient with osteoporosis occurs on the third visit.

c. Now you see the pattern. What is the probability that the first patient with osteoporosis occurs on the fourth visit? On the fifth visit? Give the general result: What is the probability that the first patient with osteoporosis occurs on the kth visit?

Comment: The distribution of the number of trials to the first “success” (in our example, a patient with osteoporosis) is called a **geometric distribution.** In this problem you have found geometric distribution probabilities when the probability of a success on each trial is 7%. The same idea works for any probability of success.

12.52 Sex selection. A couple considers continuing to have children until their first girl is born; XX is the total number of children the couple has.

We saw in Example 9.1 that, in the United States, approximately 48.8% of all newborns are baby girls. The sex of a child in successive births can be considered independent.

a. Does Xx have a binomial distribution? Explain your reasoning.
b. Obtain the probability that the couple would have only 1 child.
c. Obtain the probability that the couple would have exactly 5 children.

12.53 Overweight Americans. The National Center for Health Statistics estimates that 3% of adult American women are underweight, 42% have a healthy weight, and 55% are overweight or obese.[23] A gynecologist sees 3 patients for a routine exam.

a. What is the probability that the first patient has a healthy weight and the next two patients are overweight?
b. What is the probability that the second patient has a healthy weight and the other two patients are overweight? What is the probability that the first two patients are overweight and the last patient has a healthy weight?
c. Using the results from parts a and b, give the probability that out of any three patients, one has a healthy weight and two are overweight. This is a **multinomial probability** with three possible outcomes. Multinomial probabilities can represent any number of possible outcomes, generalizing the binomial setting (their mathematical description is beyond the scope of this introductory textbook). Binomial probabilities are a special case of the multinomial model when only two outcomes are possible.

CHAPTER 13 Sampling Distributions

mishatc/Depositphotos

IN THIS CHAPTER WE COVER...

- Parameters and statistics
- Statistical estimation and sampling distributions
- The sampling distribution of $\bar{x}$
- The central limit theorem
- The sampling distribution of $\hat{p}$
- The law of large numbers

The Centers for Disease Control and Prevention (CDC) offers a graph of cigarette use among adults in the United States, with data broken down by state.[1] The graph for 2014 indicates that 9.7% of adults in Utah were cigarette smokers—a finding that is reported along with a “low confidence limit” of 9.1% and a “high confidence limit” of 10.3%. In contrast, 26.7% of adults in West Virginia were cigarette smokers, with low and high confidence limits of 25.3% and 28.1%, respectively. These confidence limits reflect the fact that the values 9.7% and 26.7% are the percent of individuals who were cigarette smokers among a random sample of adults from each state, rather than among all adults living in these two states.

In Chapter 6 we introduced the idea of taking a random sample as a way of obtaining unbiased information about a population, when the entire population cannot be studied. This is the foundation of *statistical inference*. While the population of interest is unique, a very large number of different samples can potentially be selected. Because random samples and randomized comparative experiments include an element of chance, some variability is introduced in the results. Using the laws of probability we can ask “What would happen if we did this many times?” in order to quantify the uncertainty in sample results. This chapter presents some important answers that will be needed to work out the mathematical framework of statistical inference in the following chapters.

PARAMETERS AND STATISTICS

If we are to use sample data to draw conclusions about a wider population, we must keep straight whether a specific number describes a sample or a population. We first introduced this concept in Chapter 2. Recall the vocabulary that we used.

PARAMETER, STATISTIC

A **parameter** is a number that describes the population. In statistical practice, the value of a parameter is not known, because we cannot examine the entire population.

A **statistic** is a number that can be computed from the sample data without making use of any unknown parameters. In practice, we often use a statistic to estimate an unknown parameter.

EXAMPLE 13.1 Parameter versus statistic

The National Center for Health Statistics reports that 8% of births in 2014 were low-birth-weight infants (weighing less than 2500 grams) and that the mean age of mothers at first birth was 26.3 years.[2] These numbers were computed based on the birth certificates for all 3,988,076 births registered in the United States in 2014. They are therefore *parameters*.

mishatc/Depositphotos

The polling organization Gallup interviewed a random sample of 1023 American adults in July 2016 and found that 19% had smoked cigarettes in the past week and that smokers smoked, on average, 12 cigarettes per day.[3] Because these numbers were computed from a nationwide random sample of adults rather than from all adults in the United States, they are *statistics*. We do not know the values of the corresponding parameters in the entire population of American adults.

Remember that **s**tatistics come from **s**amples, whereas **p**arameters come from **p**opulations. Understanding the distinction between a population and a sample is essential, and the notation we use must reflect this distinction.

When dealing with quantitative data, we write μ (the Greek letter mu) for the **mean of a population.** For any given population, the value of this parameter is unique. The **mean of the sample** is the familiar $\bar{x}$, the average of the n observations in the sample. This statistic would almost certainly have a different value if we chose another sample from the same population. In most situations, the value of the parameter μ is unknown. The sample mean $\bar{x}$ from a sample or an experiment is an estimate of the mean μ of the underlying population, and we will use it as the basis of statistical inference.

Likewise, when dealing with categorical data, we use the notation p to indicate the **population proportion** of individuals of a given type, a parameter with a unique value. The **sample proportion** $\hat{p}$, is the proportion of individuals of a given type in a sample of n observations. This statistic would almost certainly have a different value if we chose another sample from the same population. The sample proportion $\hat{p}$ is an estimate of the population proportion p, whose value in most situations is unknown.

EXAMPLE 13.2 Parameter versus statistic, continued

Example 13.1 described the findings from two studies: A study of births based on all U.S. birth certificates in 2014, and a study of smoking patterns based on a random sample of American adults interviewed in

> July 2016.
>
> The cited birth numbers are parameters. The proportion of all births in 2014 that were low-birth-weight babies is p=0.08.$p = 0.08$. The mean age of all mothers giving birth to their first baby in 2014 is μ=26.3$\mu = 26.3$ years.
>
> The numbers cited in the smoking study are statistics. The proportion of adults in the sample who smoke cigarettes is p^=0.19.$\hat{p} = 0.19$. The mean amount of daily smoking among the smokers in the sample is x¯=12$\bar{x} = 12$ cigarettes.

Few studies can rely on data for an entire population. The availability of detailed birth certificates for the whole population is an exception, rather than the rule. The vast majority of studies rely on random samples, as did the Gallup survey of smoking habits, or randomized comparative experiments.

APPLY YOUR KNOWLEDGE

13.1 Nerve conduction in rats. Extensive scientific research on nerve conduction has led to the conclusion that the mean refractory period in a healthy rat is **1.3** milliseconds (ms). An experiment designed to study the effect of the pesticide DDT on nerve conduction finds that the mean refractory period of 4 rats poisoned with DDT is **1.75** ms. Is each of the boldface numbers a parameter or a statistic?

13.2 Disease surveillance. According to the CDC, there were 6153 reported cases of animal rabies in the United States in all of 2010. Raccoons were the most common rabid animals, making up **36.5%** of all reported cases that year. Gallup contacted by phone a random sample of 27,300 American adults in December 2012 and found that **3.2%** had been sick with the flu. Is each of the boldface numbers a parameter or a statistic?

STATISTICAL ESTIMATION AND SAMPLING DISTRIBUTIONS

Statistical inference uses sample data to draw conclusions about the entire population. *Because good samples are chosen randomly, statistics such as* $\bar{x}$ *and* $\hat{p}$ *are random variables.* Therefore, we can describe the behavior of a sample statistic with a probability model that answers the question “What would happen if we did this many times?”

Taking many random samples from a real population is time consuming and expensive. It is much easier to obtain such samples with a computer **simulation.**

EXAMPLE 13.3 Simulated random samples

IQ tests measure some forms of human intelligence. The distribution of IQ scores in a homogeneous population (for example, the same age group) is well described by a Normal distribution. The results of an IQ test are scored on a scale designed so that the mean μ of the population of interest is 100 and the standard deviation σ is 15. Figure 13.1(a) shows the $N(100, 15)$ density curve.

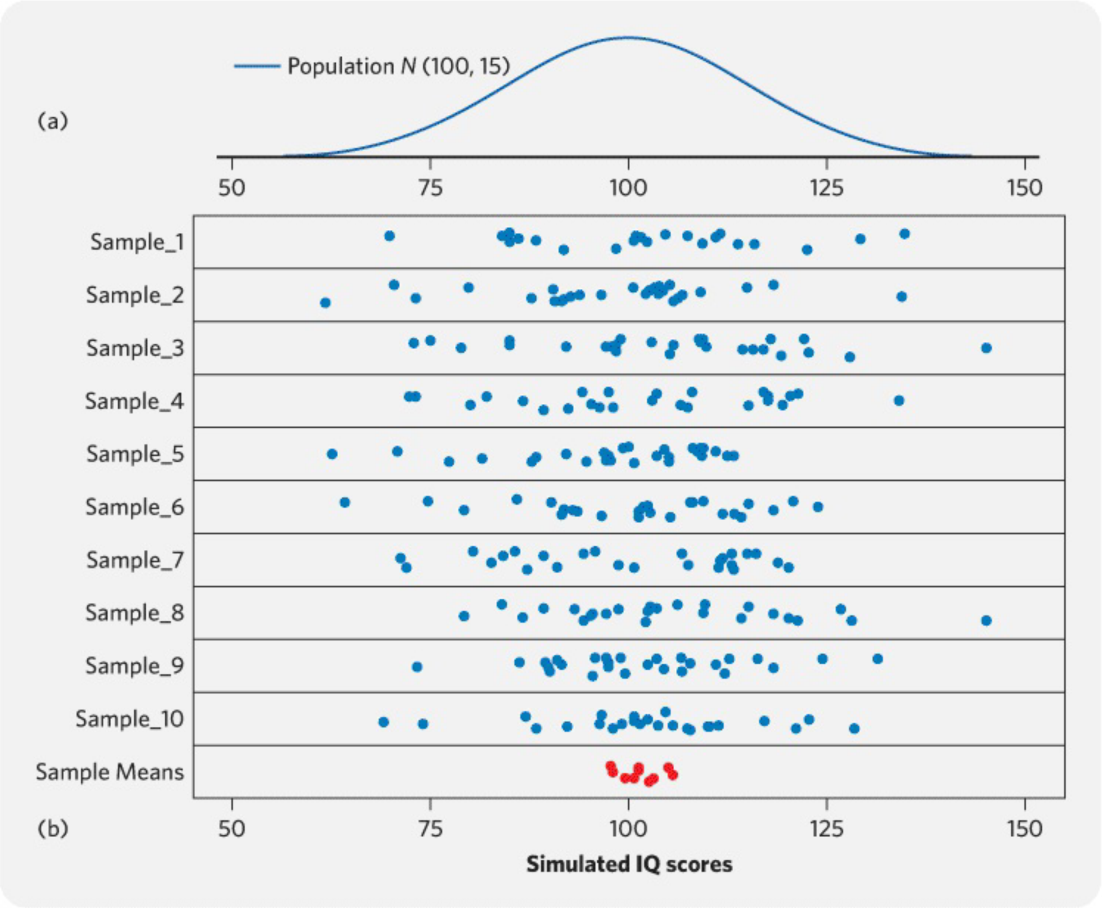
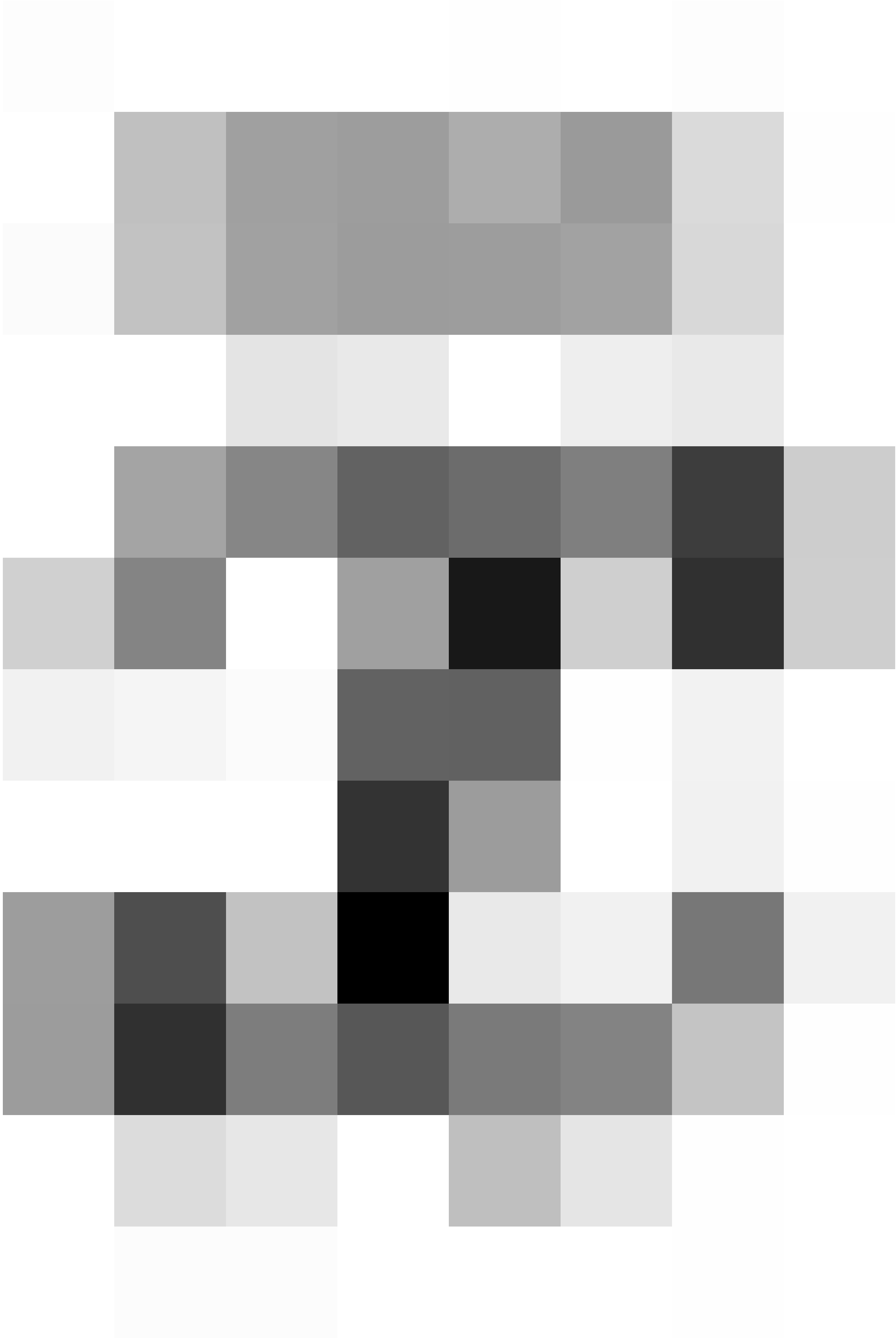


Figure 13.1

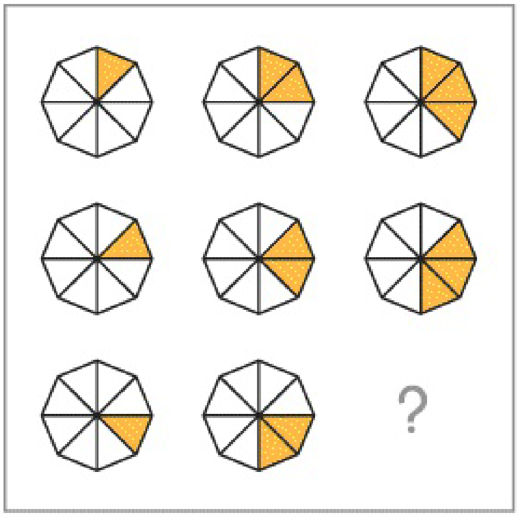
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 13.1 Simulating random samples from the $N(100, 15)$ distribution of IQ scores. (a) The population distribution $N(100, 15)$ is displayed graphically. (b) Each of the first 10 rows represents one simulated random sample of 25 individuals from the $N(100, 15)$ population. The last row shows the means $\bar{x}$



of these 10 simulated random samples.

Consider what would happen if we took an SRS of, say, 25 individuals from a given population and had them take an IQ test. The 68–95–99.7 rule for Normal curves tells us that any given IQ score under the N(100, 15) distribution has an approximately 95% chance of being a value between 70 and 130 (within two standard deviations of the mean) and only a 0.3% chance of being a value outside the range 55 to 145 (more than three standard deviations away from the mean). Figure 13.1(b) shows the dotplots of 10 different simulated random samples of 25 IQ scores.



Lyudmyla Kharlamova/Shutterstock.com

For each simulated SRS, we can compute the mean IQ score x¯.$\bar{x}$. We don’t expect any particular x¯ $\bar{x}$ to be exactly equal to μ.μ. However, because an SRS should fairly represent the population, the mean x¯ $\bar{x}$ of the sample should be somewhere near the mean μμ of the population. The last row of Figure 13.1(b) shows that the 10 mean IQ scores from the simulated samples are not identical (demonstrating that the statistic x¯ $\bar{x}$ is indeed a random variable), but they are closely clustered around the mean μ=100$\mu = 100$ of the population of IQ scores.

Example 13.2 suggests that, while the mean x¯ $\bar{x}$ of an SRS would not be a perfect guess for the unknown value of a population mean μ,μ, it might be a reasonable estimate. To better understand the properties of the sample mean x¯ $\bar{x}$ we ask, “What would happen if we took a great many samples of 25 subjects from this population?”

EXAMPLE 13.4 What would happen in many samples?

The computer simulations in Example 13.3 were created with the statistical software Minitab, using its “Random data” tool to generate sets of 25 observations from the *N*(100, 15) distribution at random. The tool can be used to simulate as many such SRSs as we want. Figure 13.2 illustrates the process.

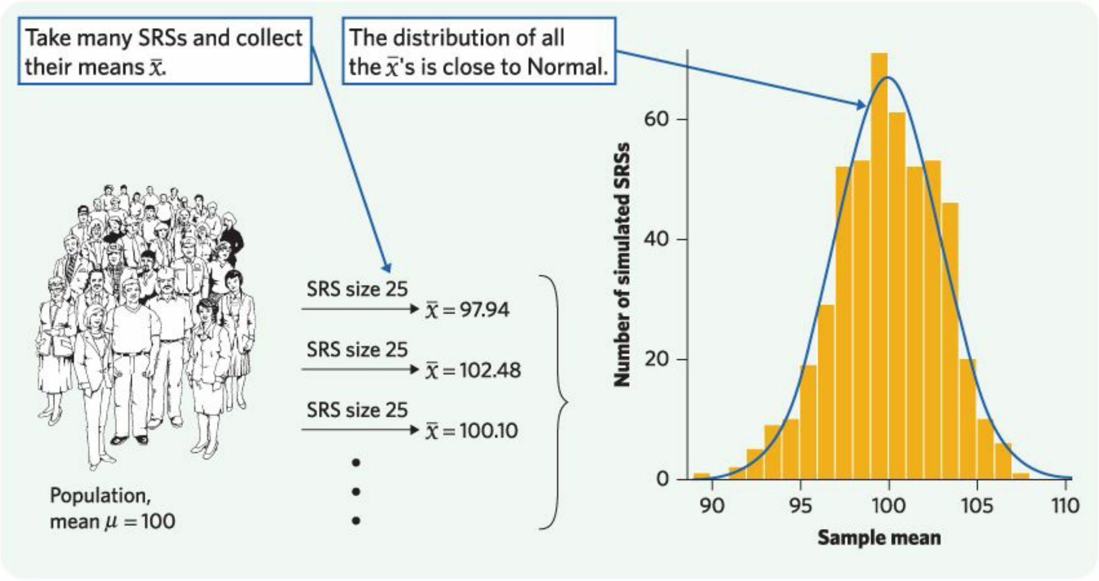
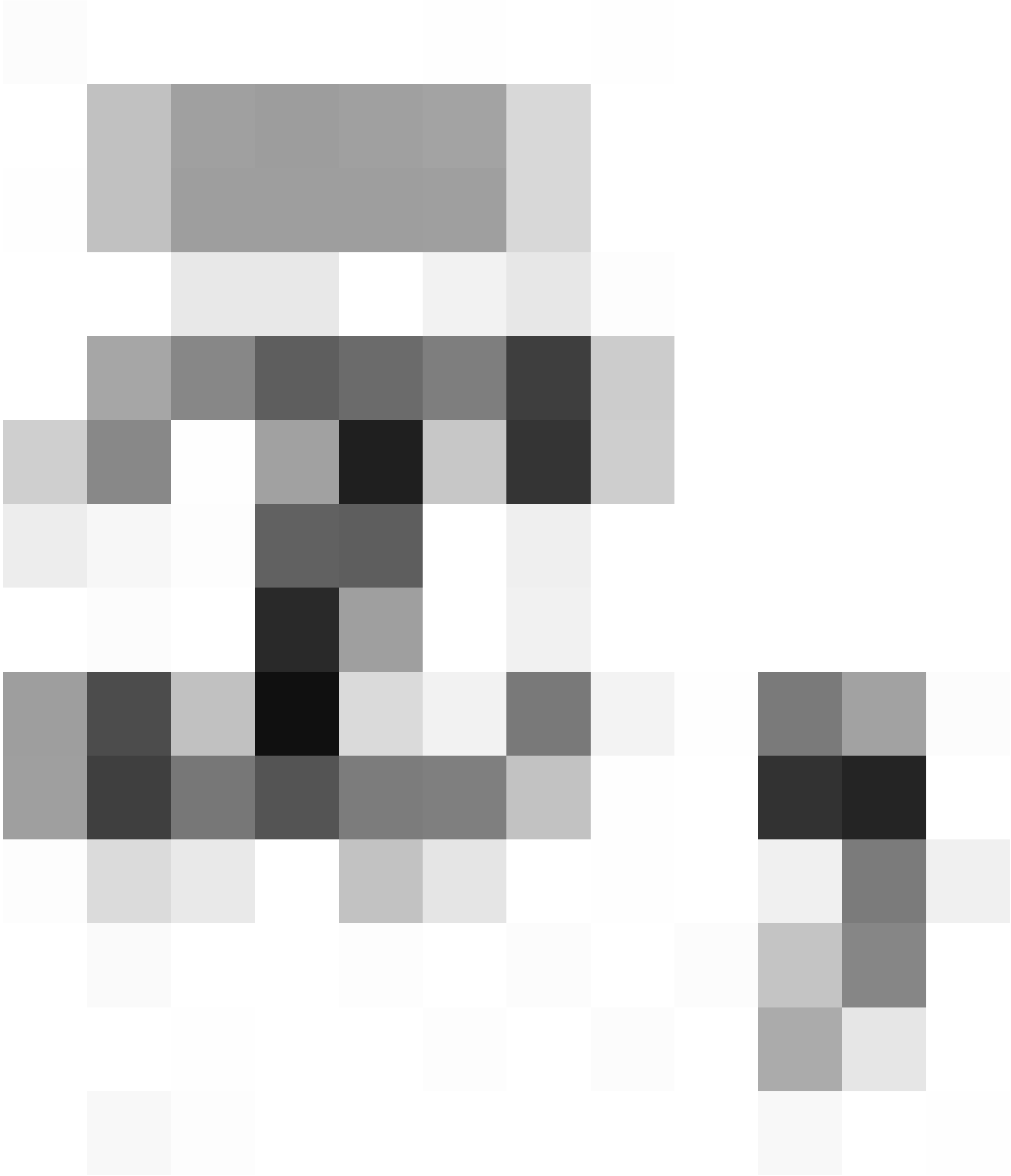
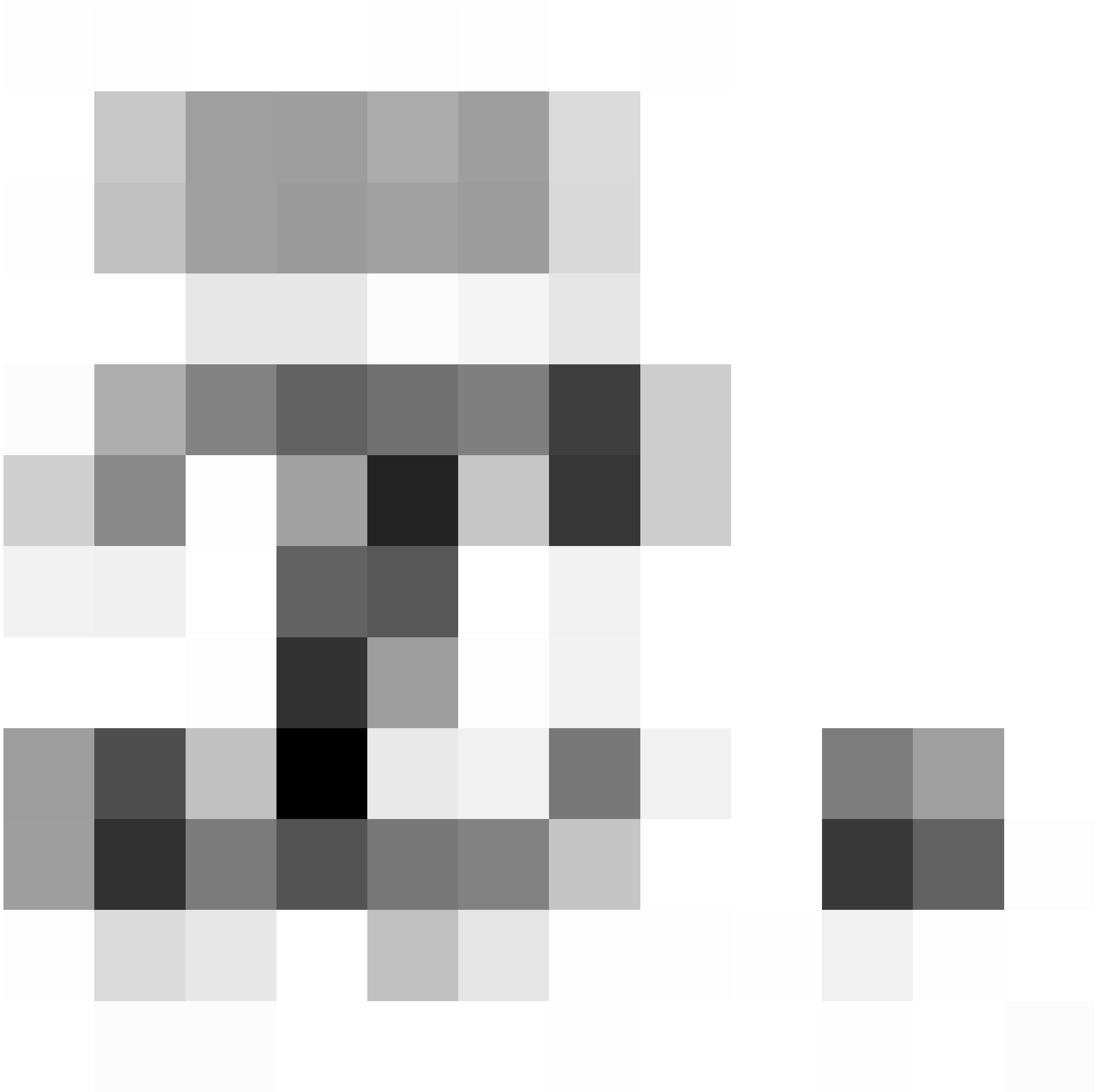


Figure 13.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 13.2 The idea of a sampling distribution: Take many SRSs from the same population, compute their sample means $\bar{x}$,



and display the distribution of $\bar{x}$.



The histogram shows the result of 500 simulated SRSs from the *N*(100, 15) population distribution of IQ scores.

Follow the flow of the figure from the population at the left, to choosing an SRS and finding the mean x¯ $\bar{x}$ for this sample, to collecting and plotting the x¯ $\bar{x}$ values for many samples at the right. The first sample has x¯=97.94.$\bar{x} = 97.94.$ The second sample contains a different set of 25 observations, with x¯=102.48,$\bar{x} = 102.48,$ and so on. The histogram at the right of the figure shows the distribution of the values of x¯ $\bar{x}$ from 500 separate SRSs of size 25. This histogram displays the underlying structure of the *sampling distribution* of the statistic x¯.$\bar{x}.$

SAMPLING DISTRIBUTION

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

The sampling distribution is the ideal pattern that would emerge if we looked at all possible samples of size 25 from our population. A distribution obtained from a fixed number of SRSs, like the 500 SRSs in Figure 13.2, is merely an approximation of the sampling distribution. Such an approximation is reasonably good only if we collect or simulate a large number of SRSs. One of the uses of probability theory in statistics is to obtain exact sampling distributions without simulation. The interpretation of a sampling distribution is the same, however, whether we obtain it by simulation or by the mathematics of probability.

We can use the tools of data analysis to describe any distribution. Let's apply those tools to Figure 13.2. What can we say about the shape, center, and spread of this distribution?

- **Shape:** It looks Normal! The Normal curve superimposed onto the histogram is a very good model for the distribution of $\bar{x}$ from the 500 SRSs.
- **Center:** The mean of these 500 $\bar{x}$ values is 99.96. That is, the distribution is centered very close to the population mean $\mu=100$.
- **Spread:** The standard deviation of these 500 $\bar{x}$ values is 2.99, notably smaller than the standard deviation $\sigma=15$ of the population of individual subjects.

In Example 13.3 we studied the behavior of $\bar{x}$ when we simulated random samples of $n=25$ observations. Would the findings be substantially different if we chose a different sample size n? Figure 13.3 expands on Figure 13.1. It displays the distribution of the sample means for 10 simulated SRSs when the sample size n is either 4, 25, 64, or 400. We see clearly that the spread of the sample means is smaller when the means are computed from larger samples, yet the center remains the same.

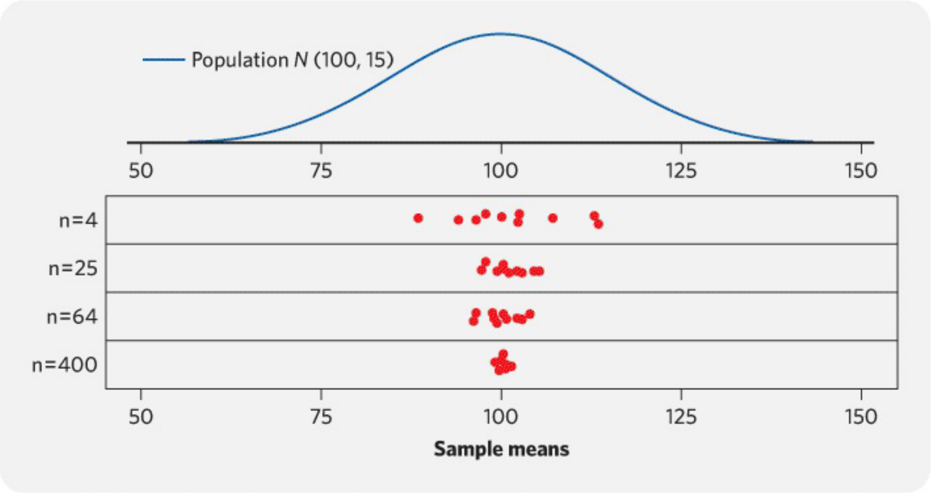
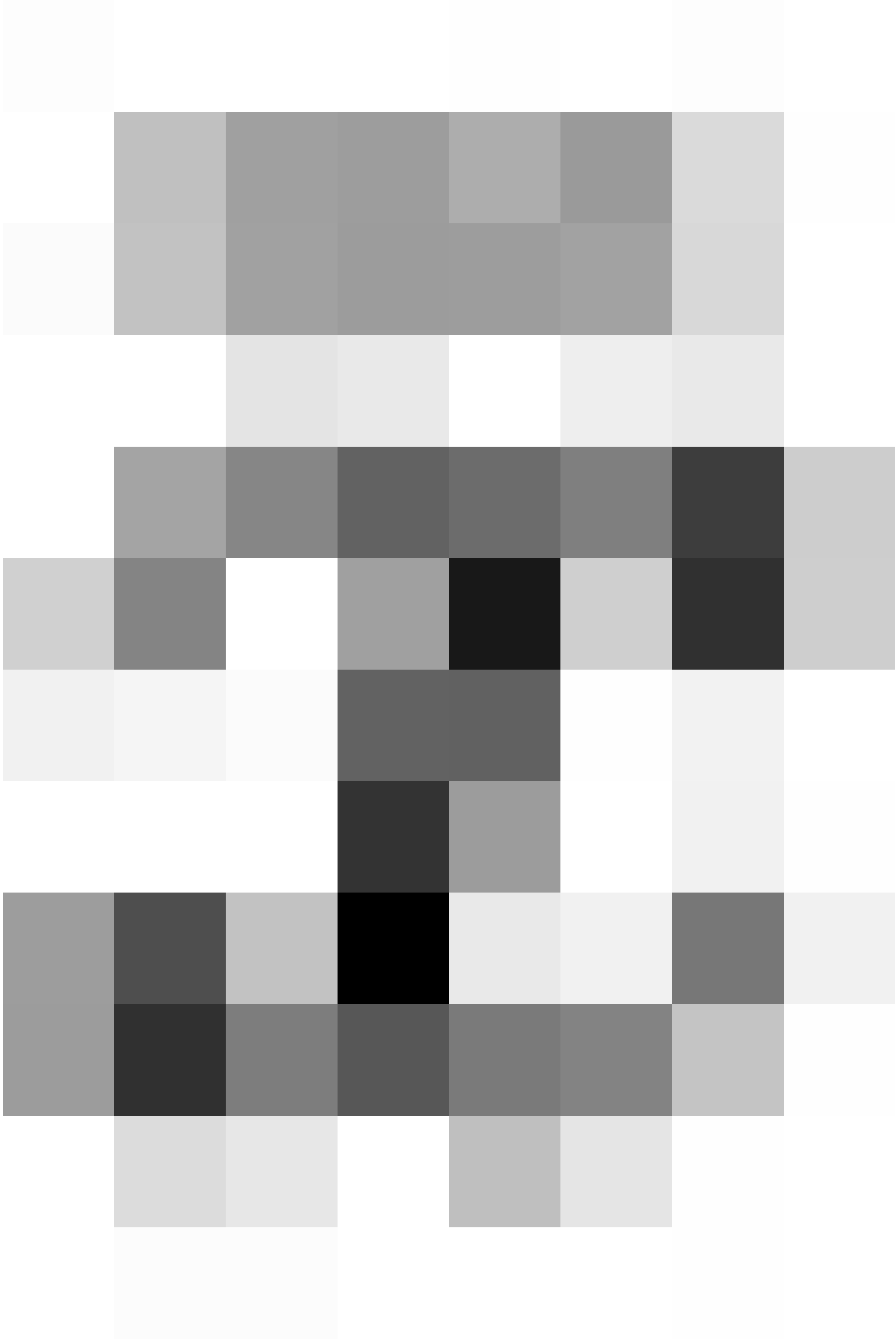


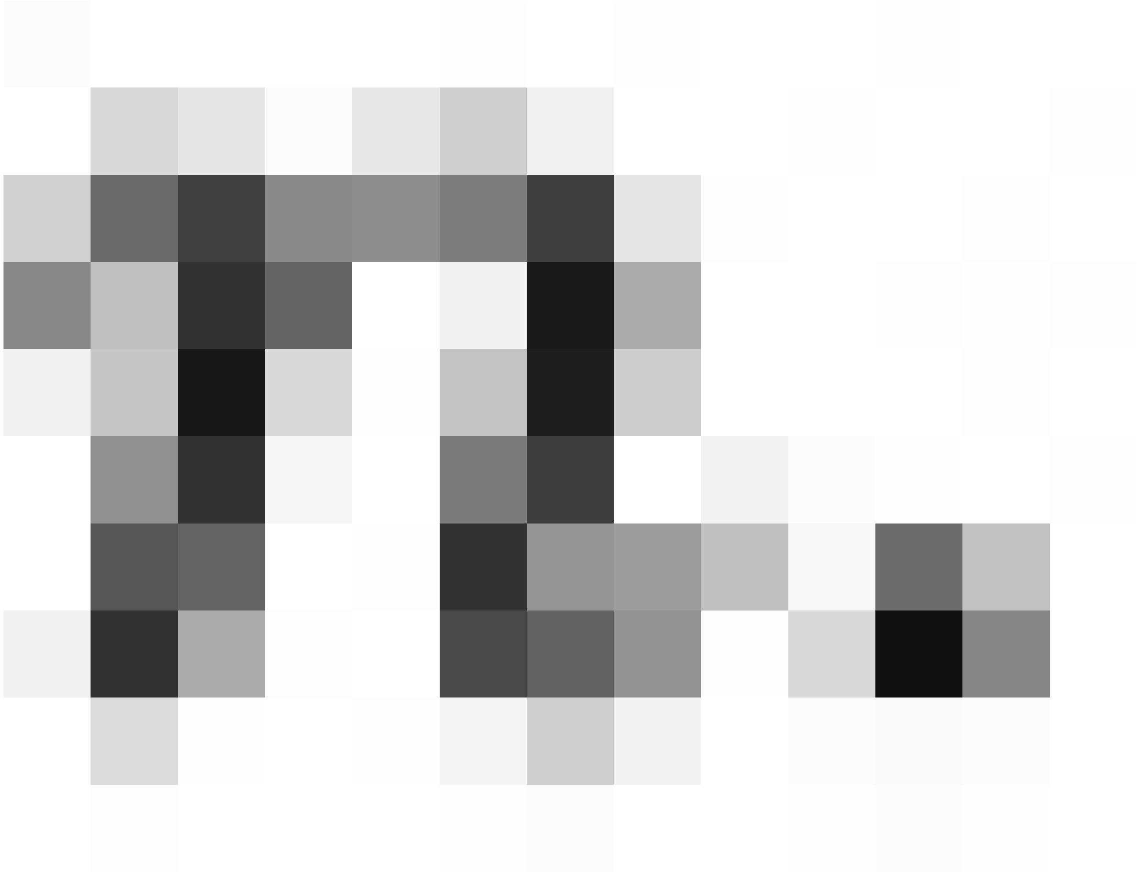
Figure 13.3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 13.3 Population distribution $N(100, 15)$ at the top. Each row underneath it shows the means $\bar{x}$



of 10 simulated SRSs in each of four situations varying by the chosen sample size n.



Although these results describe simulations of a sampling distribution, they reflect facts that are true whenever we use random sampling. We shall describe these properties next.

APPLY YOUR KNOWLEDGE

13.3 Generating a sampling distribution for pedagogical purposes. Let’s illustrate the idea of a sampling distribution in the case of a very small sample from a very small population. The population is the 10 female students in a class:

Student	0	1	2	3	4	5	6	7	8	9
Weight	136	99	118	129	125	170	130	128	120	147

The parameter of interest is the mean weight in pounds μ in this population. The sample is an SRS of size $n=4$ drawn from the population. Because the students are labeled 0 to 9, a single random digit from Table A (at the back of the book) chooses one student for the sample.

a. Find the mean of the 10 weights in the population. This is the population mean μ.μ.
b. Use the first four digits in row 116 of Table A to draw an SRS of size 4 from this population. What are the four weights in your sample? What is their mean x¯?$\bar{x}$? This statistic is an estimate of μ.μ.
c. Repeat this process nine more times, using the first four digits in rows 117 to 125 of Table A. Make a histogram of the 10 values of x¯.$\bar{x}$. You are constructing the sampling distribution of x¯.$\bar{x}$. Is the center of your histogram close to μ?μ?
d. This is a lengthy process, even for such a small population. Explain why it would be unrealistic to attempt to build a sampling distribution by hand for a large population.

13.4 A sampling distribution. We can use the *Simple Random Sample* applet to help us grasp the idea of a sampling distribution (for pedagogical purposes). In the applet, form a population labeled 1 to 100. We will choose an SRS of 10 of these numbers. That is, in this exercise, the numbers themselves are the population, not just labels for 100 individuals. The mean of the whole numbers 1 to 100 is μ=50.5.$\mu = 50.5$. This is the population mean.

a. Use the applet to choose an SRS of size 10. Which 10 numbers were chosen? What is their mean? This is the sample mean x¯.$\bar{x}$.
b. Although the population and its mean μ=50.5$\mu = 50.5$ remain fixed, the sample mean changes as we take more samples. Take another SRS of size 10. (Use the “Reset” button to return to the original population before taking the second sample.) What are the 10 numbers in your sample? What is their mean? This is another value of x¯.$\bar{x}$.
c. Take eight more SRSs from this same population and record their means. You now have 10 values of the sample mean x¯$\bar{x}$ from 10 SRSs of the same size from the same population. Make

a histogram of the 10 values and mark the population mean $\mu=50.5$ on the horizontal axis. Are your 10 sample values roughly centered at the population value μ? (If you kept going forever, your $\bar{x}$-values would form the sampling distribution of the sample mean; the population mean μ would indeed be the center of this distribution.)

d. Explain some of the obstacles to trying to build a sampling distribution through repeated samples, even with the help of software.

THE SAMPLING DISTRIBUTION OF $\bar{x}$

Figure 13.2 suggests that when we choose many SRSs from a population, the sampling distribution of the sample means is centered at the mean of the original population and is less spread out than the distribution of individual observations. Although we won't elaborate on this point, it can be demonstrated mathematically. Here are just the facts.

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

Suppose that $\bar{x}$ is the mean of an SRS of size n drawn from a large population with mean μ and standard deviation σ. Then the sampling distribution of the statistic $\bar{x}$ has **mean** μ and **standard deviation** $\sigma/\sqrt{n}$.

These facts about the mean and the standard deviation of the sampling distribution of $\bar{x}$ are true for *any* population, provided that the population is much larger than the sample (say, at least 20 times larger).[4] Both facts have important implications for statistical inference.

- The mean of the statistic $\bar{x}$ is always equal to the mean μ of the population. That is, the sampling distribution of $\bar{x}$ is centered at μ. In repeated sampling, $\bar{x}$ will sometimes fall above the true value of the parameter μ and sometimes below it, but there is no systematic tendency to overestimate or underestimate the parameter. This makes the idea of lack of bias in the sense of "no favoritism" more precise. Because the mean of all $\bar{x}$'s in the sampling distribution is equal to μ, we say that the statistic $\bar{x}$ is an **unbiased estimator** of the parameter μ.
- An unbiased estimator is "correct on average" over many samples. How close the estimator is to the parameter in most samples is determined by the spread of the sampling distribution. If individual observations ($n = 1$) have standard deviation σ, then sample means $\bar{x}$ from samples of size n (when $n > 1$) have standard

deviation $\sigma/\sqrt{n}$. That is, **averages are less variable than individual observations,** and averages computed from larger samples are less variable than averages computed from smaller samples.

We have described the center and spread of the sampling distribution of the sample mean $\bar{x}$, but not its shape. The shape of the distribution of $\bar{x}$ depends on the shape of the population. Here is one important case: If measurements in the population follow a Normal distribution, then so does the sample mean.

SAMPLING DISTRIBUTION OF A SAMPLE MEAN FOR A NORMAL POPULATION

If individual observations have the $N(\mu, \sigma)$ distribution, then the sampling distribution of the mean $\bar{x}$ of an SRS of size n has the $N(\mu, \sigma/\sqrt{n})$ distribution.

EXAMPLE 13.5 Population distribution versus sampling distribution

The IQ scores of individuals in a given age group follow the Normal distribution, with mean $\mu = 100$ and standard deviation $\sigma = 15$. We call this the **population distribution** because it shows how measurements vary within the population.

Figure 13.2 shows the means $\bar{x}$ of many SRSs of size 25 from this population. The *sampling distribution* describes how the values of $\bar{x}$ vary among all possible samples of size 25 from this population. That sampling distribution is also Normal, with mean $\mu = 25$ and standard deviation

$$\frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = 3$$

Figure 13.4 contrasts these two Normal distributions. Both are centered at the population mean $(\mu = 100)$, but sample means are much less variable than individual observations. Based on the 68–95–

99.7 rule (page 274), we expect roughly 95% of all individual IQ scores to be values between 70 and 130 (within two standard deviations of the mean) but we expect roughly 95% of all mean IQ scores for random samples of 25 individuals to be values between 94 and 106.

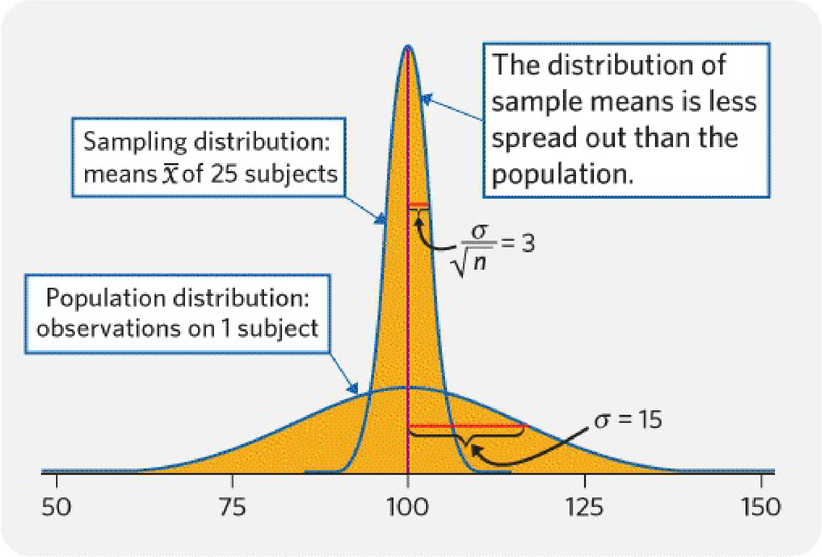
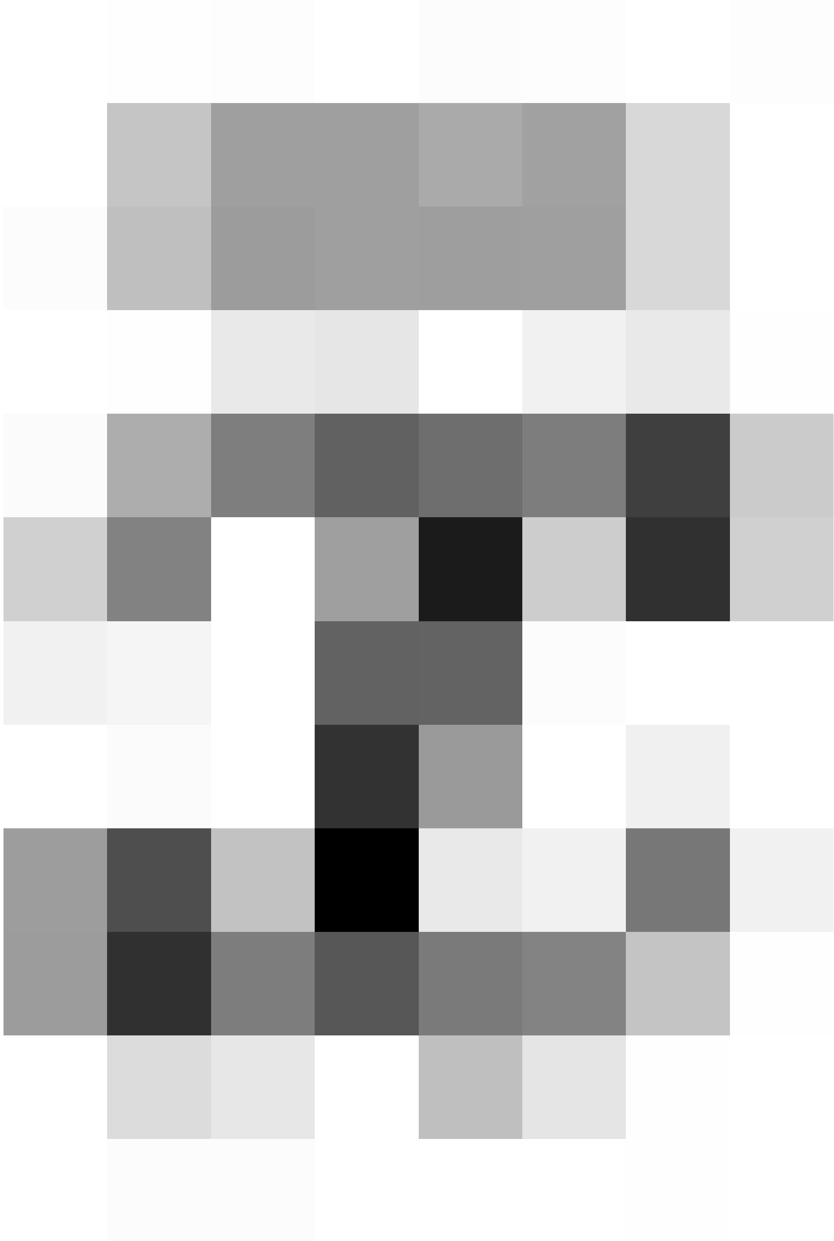


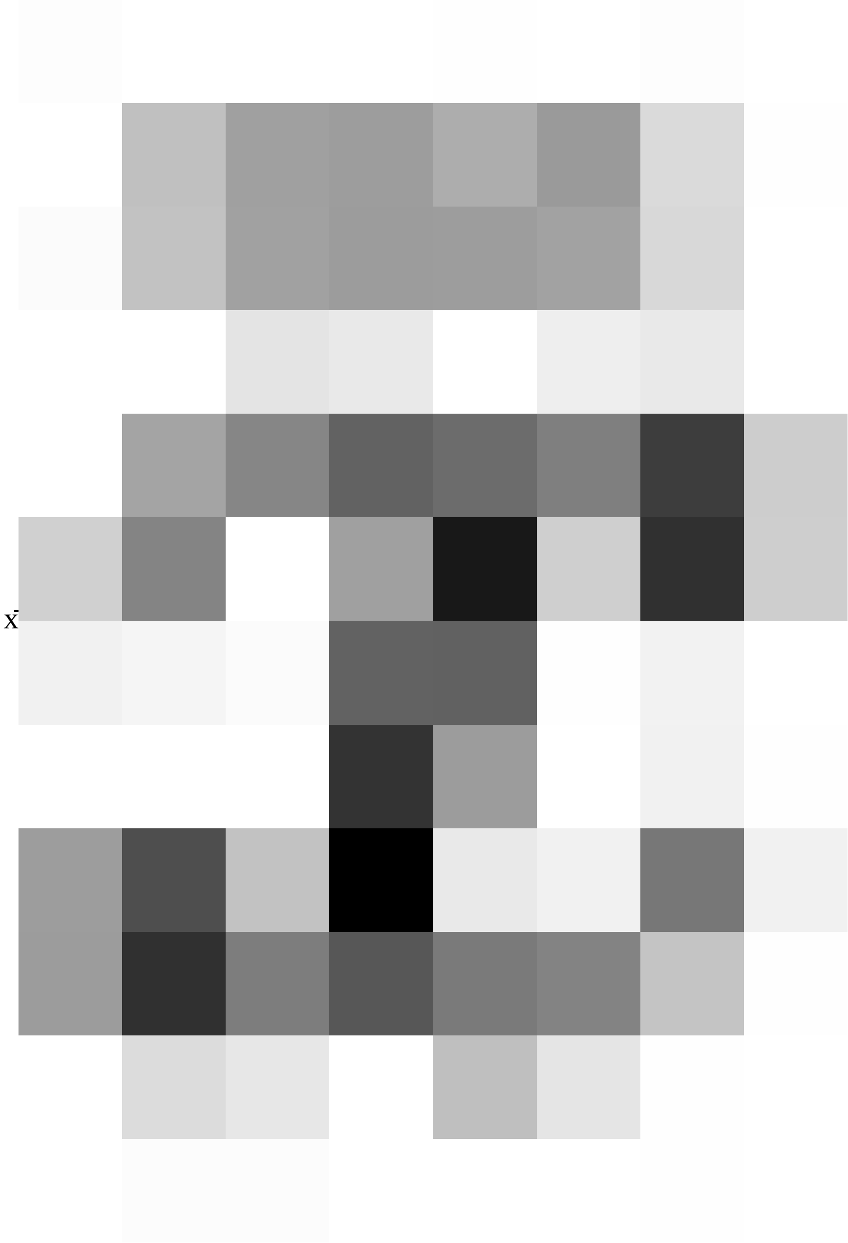
Figure 13.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Companyt

FIGURE 13.4 The distribution of single observations (the population) compared with the distribution of the means $\bar{x}$



of 25 observations

(the sampling distribution). Because the population has a Normal distribution, the sampling



distribution of $\bar{x}$ is also Normal, but it is narrower. Averages are less variable than individual observations.

Not only is the standard deviation of the sampling distribution of $\bar{x}$ smaller than the standard deviation of individual observations, but it gets smaller as we take larger samples. **The results of large random samples are less variable than the results of small random samples.** If n is large, the standard deviation of $\bar{x}$ is small, and almost all random samples will give values of $\bar{x}$ that lie very close to the true parameter μ. That is, the sample mean from a large random sample can be trusted to estimate the population mean accurately. This property of sampling distributions has many important, direct applications in statistics. *However, the standard deviation of the sampling distribution gets smaller only at the rate* $\sqrt{n}$. *To cut the standard deviation of* $\bar{x}$ *by 10, we must take 100 times as many observations—not just 10 times as many—and large sample sizes are not always an option.*

APPLY YOUR KNOWLEDGE

13.5 A sample of young men. A government sample survey plans to measure the blood cholesterol levels of an SRS of men age 20 to 34 years. The researchers will report the mean $\bar{x}$ from their sample as an estimate of the mean cholesterol level μ in this population.

a. Explain to someone who knows nothing about statistics what it means to say that $\bar{x}$ is an “unbiased” estimator of μ.

b. The sample result $\bar{x}$ is an unbiased estimator of the true population μ no matter what size SRS the study uses. Explain to someone who knows nothing about statistics why a large sample gives more trustworthy results than a small sample.

13.6 Measurements in the lab. Juan makes a measurement in a chemistry laboratory and records the result in his lab report. The standard deviation of lab measurements made by students is $\sigma = 10$ milligrams. Juan repeats the measurement 3 times and records the mean $\bar{x}$ of his 3 measurements.

a. What is the standard deviation of Juan’s mean result? (That is, if Juan kept making sets of 3 measurements and averaging them, what would be the standard deviation of all his $\bar{x}$’s?)

b. How many times must Juan repeat the measurement to reduce the standard deviation of $\bar{x}$ to 5? Explain to someone who knows nothing about statistics the advantage of reporting the average of several measurements rather than the result of a single measurement.

© Spencer Grant/PhotoEdit

13.7 A sample of young men, continued. Suppose that the blood cholesterol levels of all men age 20 to 34 years follows the Normal distribution with mean $\mu = 188$ milligrams per deciliter (mg/dl) and standard deviation $\sigma = 41$ mg/dl.

a. Choose an SRS of 100 men from this population. What is the sampling distribution of $\bar{x}$? What is the probability that $\bar{x}$ takes a value between 185 and 191 mg/dl? This is the probability that $\bar{x}$ estimates μ within ± 3 mg/dl.
b. Choose an SRS of 1000 men from this population. Now what is the probability that $\bar{x}$ falls within ± 3 mg/dl of μ? The larger sample is much more likely to give an accurate estimate of μ.

THE CENTRAL LIMIT THEOREM

The facts about the mean and standard deviation of $\bar{x}$ are true no matter what the shape of the population distribution may be. Figure 13.5(a) shows the density curve of a strongly skewed distribution called an *exponential distribution.* Exponential distributions are good models of, for example, radioactive decay and the time the body takes to metabolize and excrete drugs. The mean μ of the exponential distribution displayed is 1, and its standard deviation σ is also 1. Figure 13.6(a) shows the Normal density curve with mean $\mu = 100$ and standard deviation $\sigma = 15$ representing the distribution of IQ scores in a population of a given age group described in Example 13.3. In both figures, parts b, c, and d show, respectively, the density curves of the sample means of 2, 10, and 25 observations from these populations. We can see that, whether the population has a skewed exponential distribution or a symmetric Normal distribution, the sampling distributions of $\bar{x}$ are all centered on the value of the population mean μ, and the sampling distributions for the larger values of n have the smaller spread (reflecting the smaller value of $\frac{\sigma}{\sqrt{n}}$).

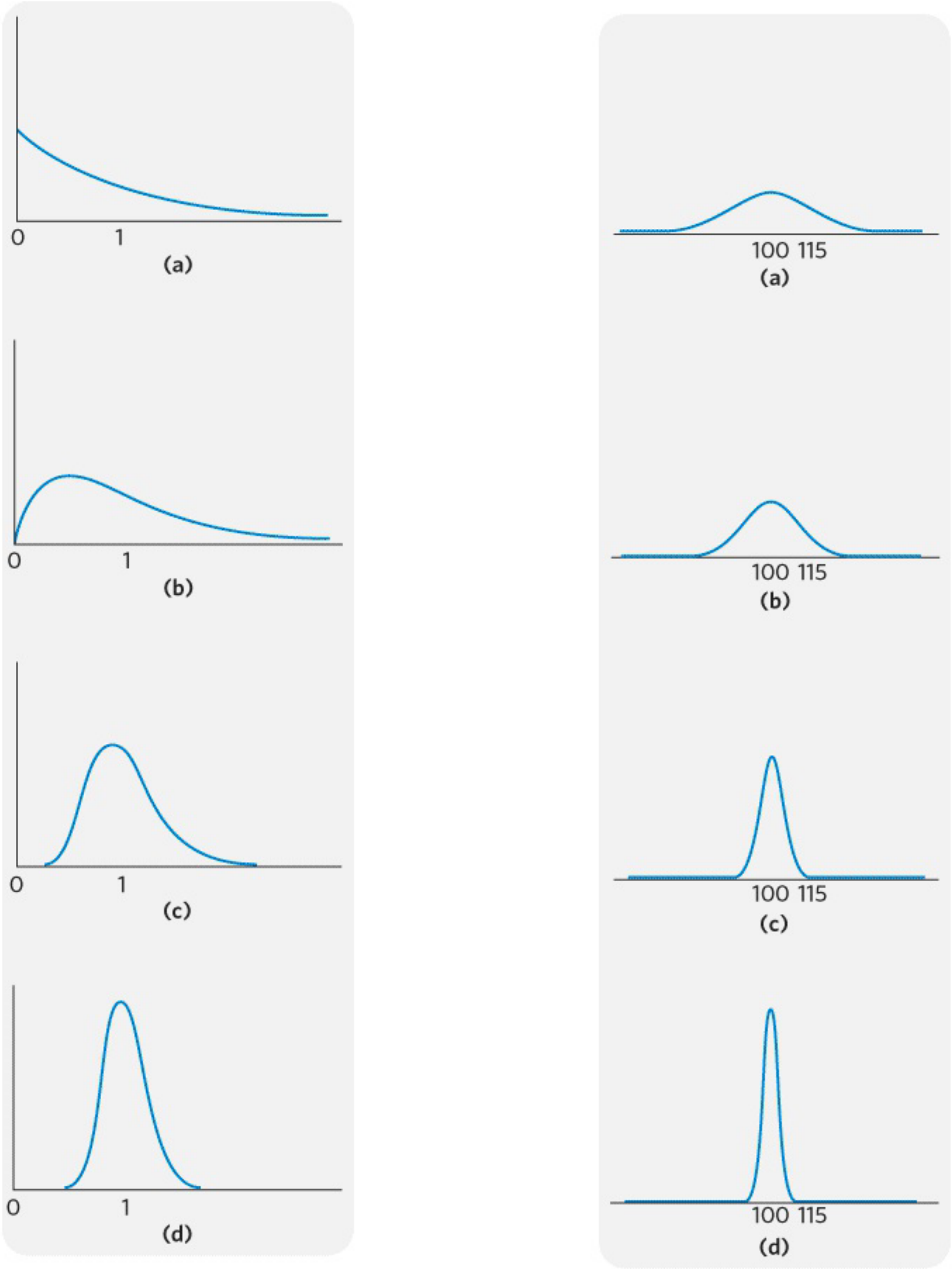
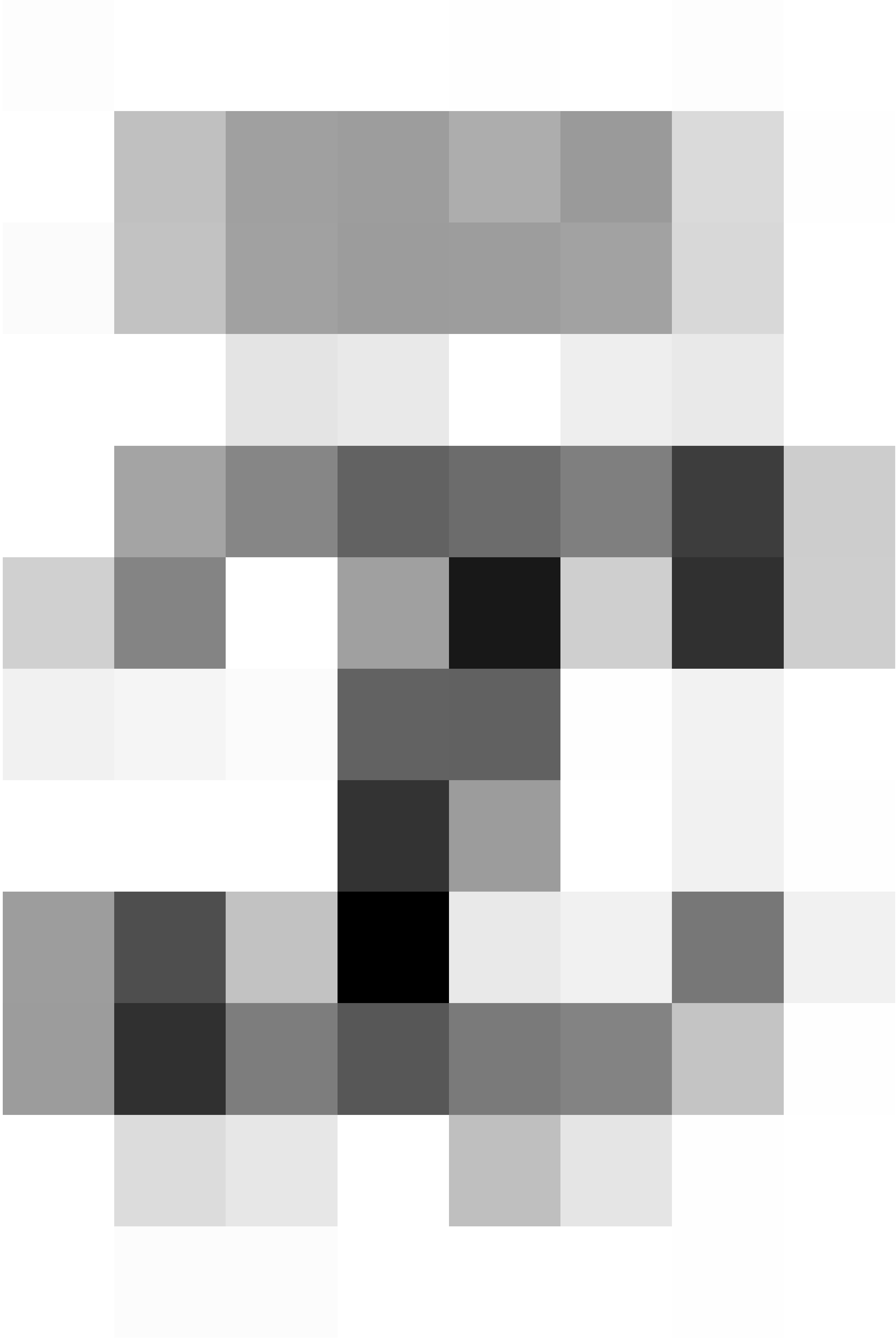


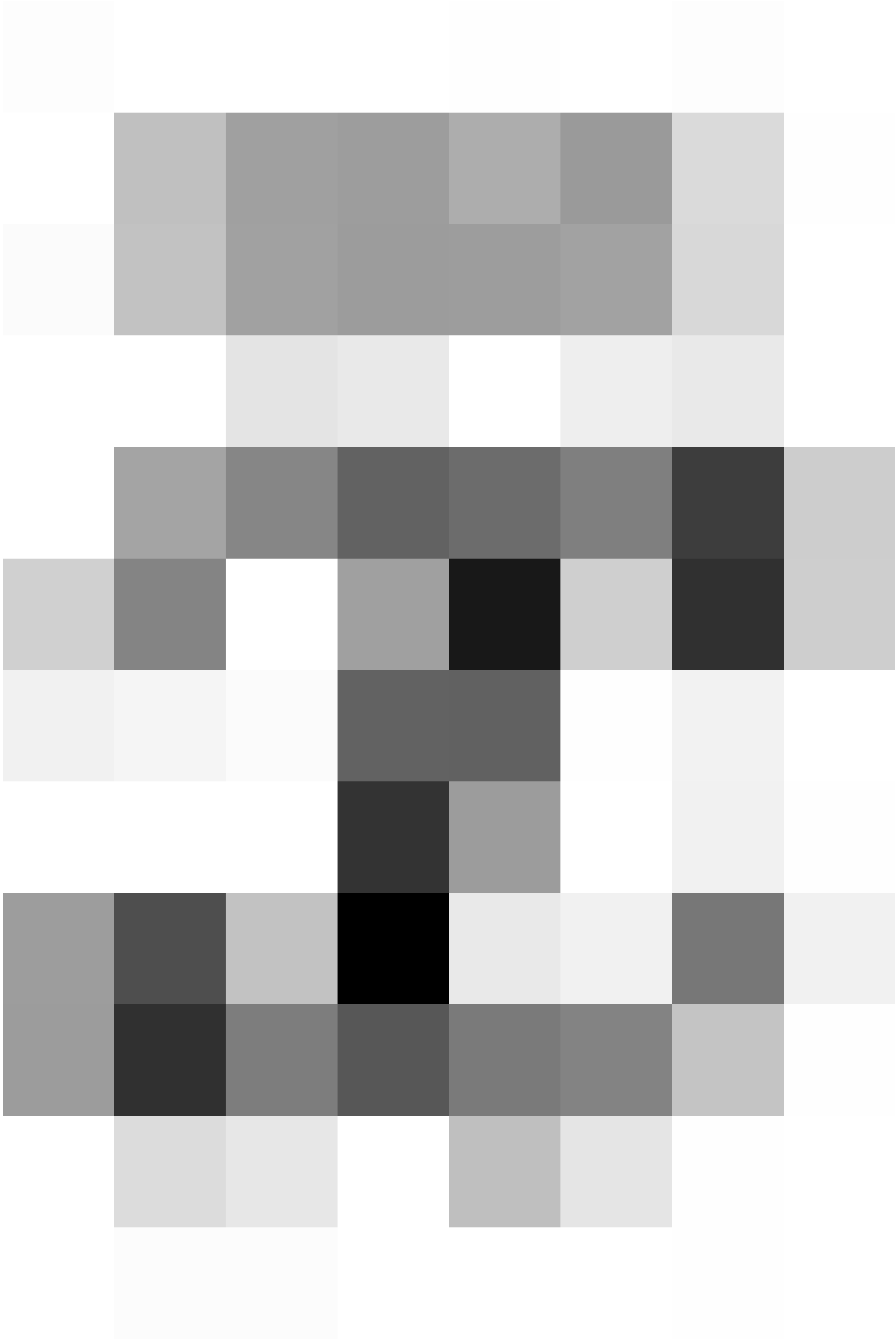
Figure 13.5 (Left), Figure 13.6 (Right)

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

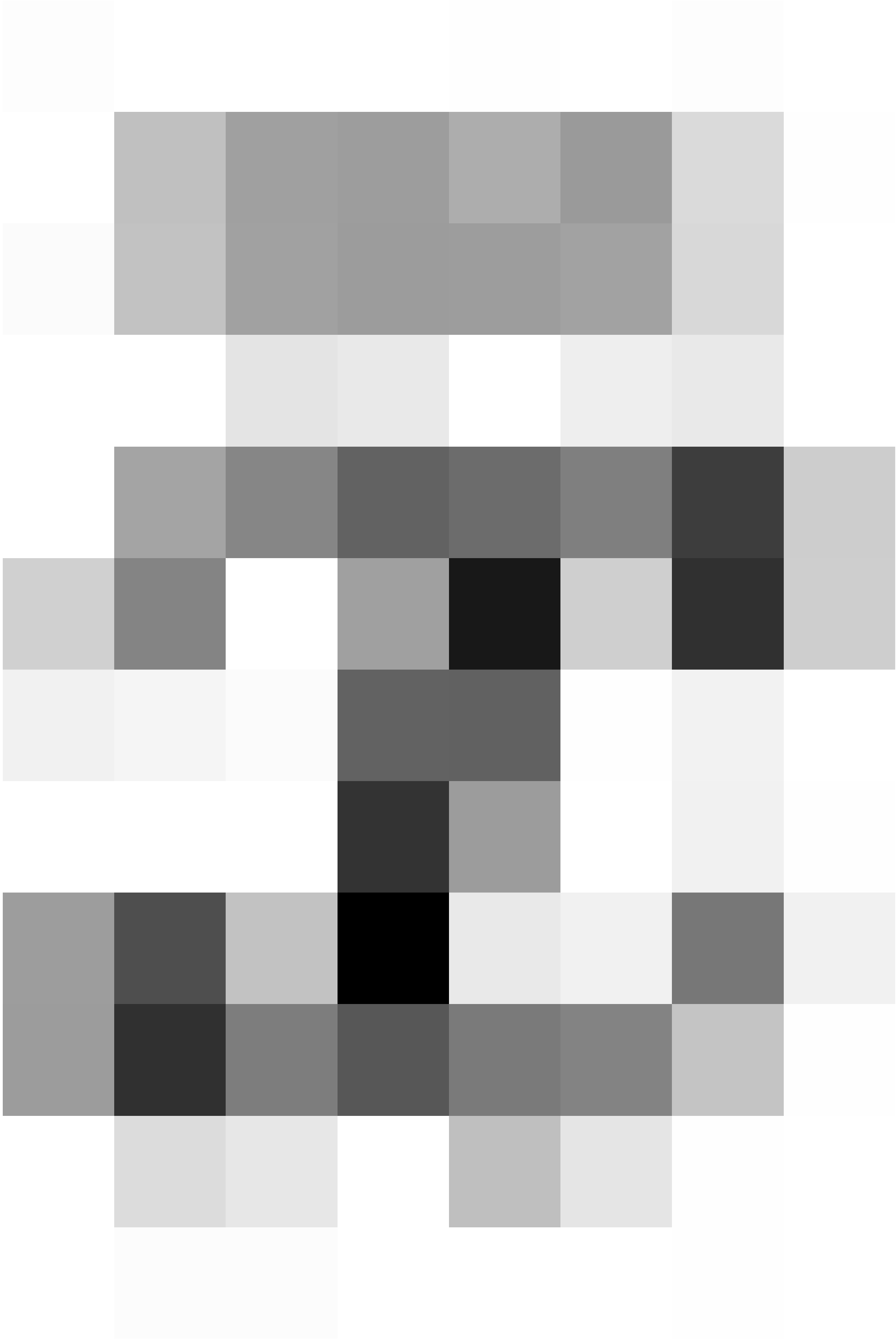
FIGURE 13.5 **(left)** The central limit theorem in action: The distribution of sample means $\bar{x}$



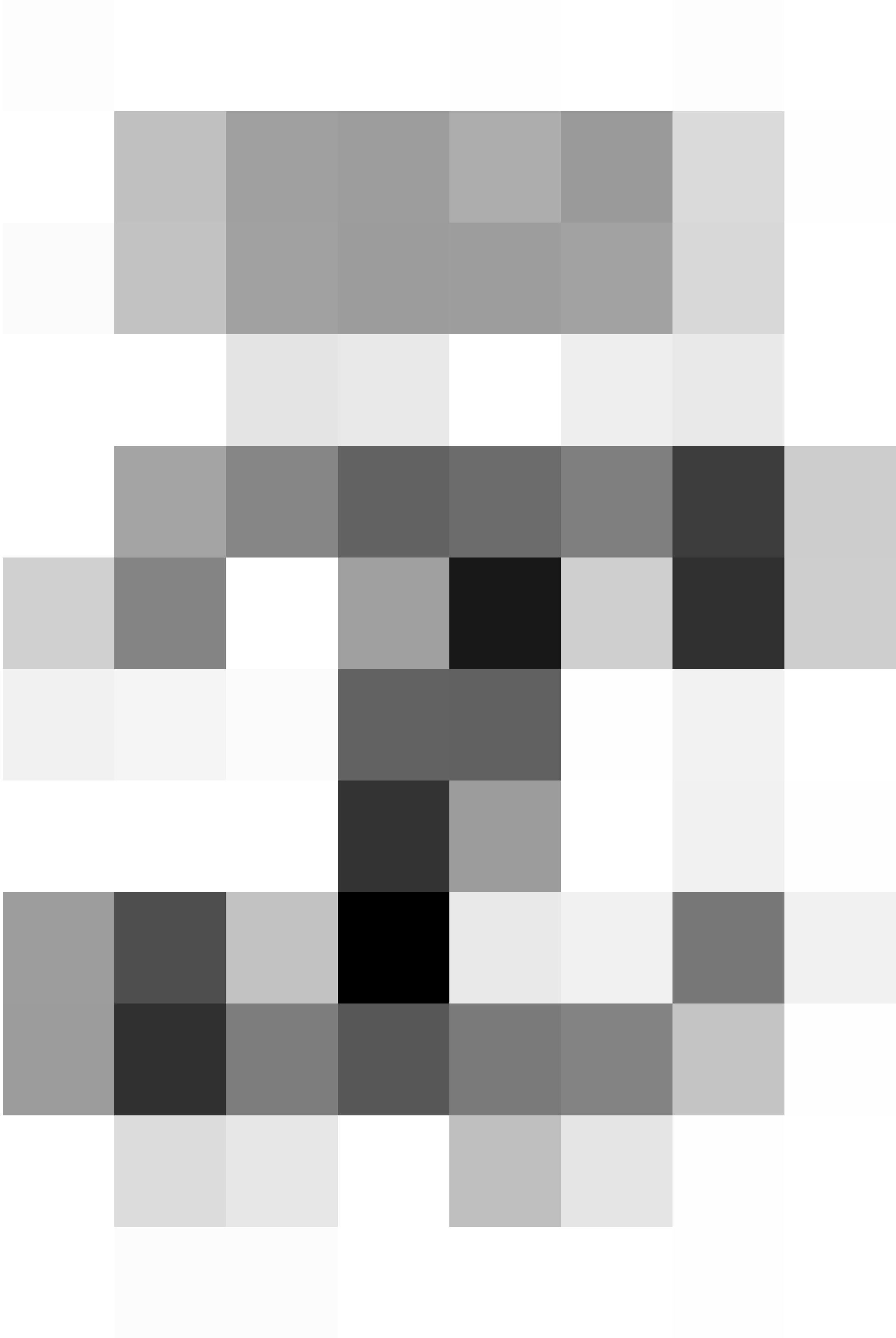
from a strongly non-Normal population is more Normal for larger sample sizes. (a) The distribution of single observations (the population distribution). (b) The sampling distribution of $\bar{x}$



for samples of 2 observations. (c) The sampling distribution of $\bar{x}$

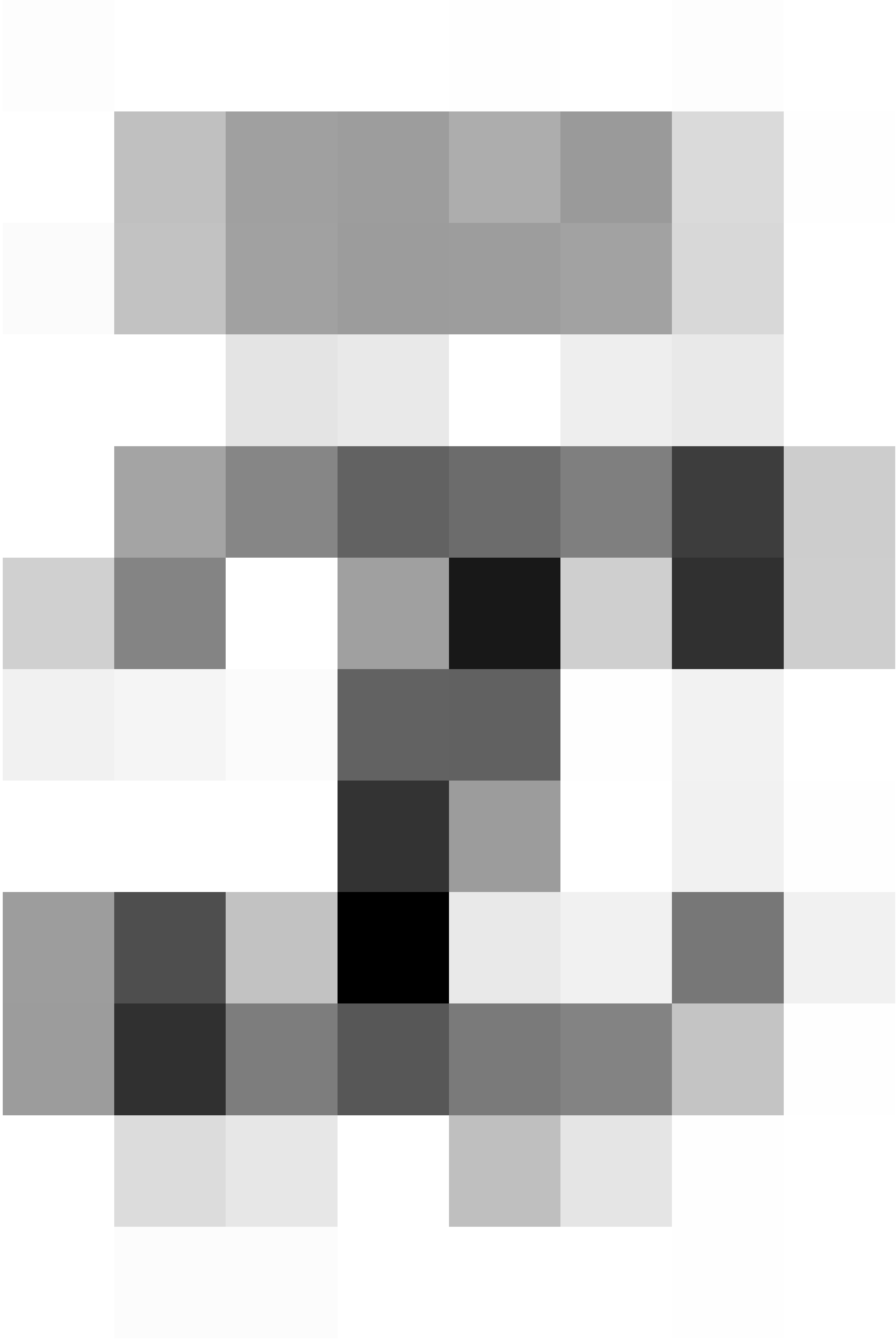


for samples of 10 observations. (d) The sampling distribution of $\bar{x}$

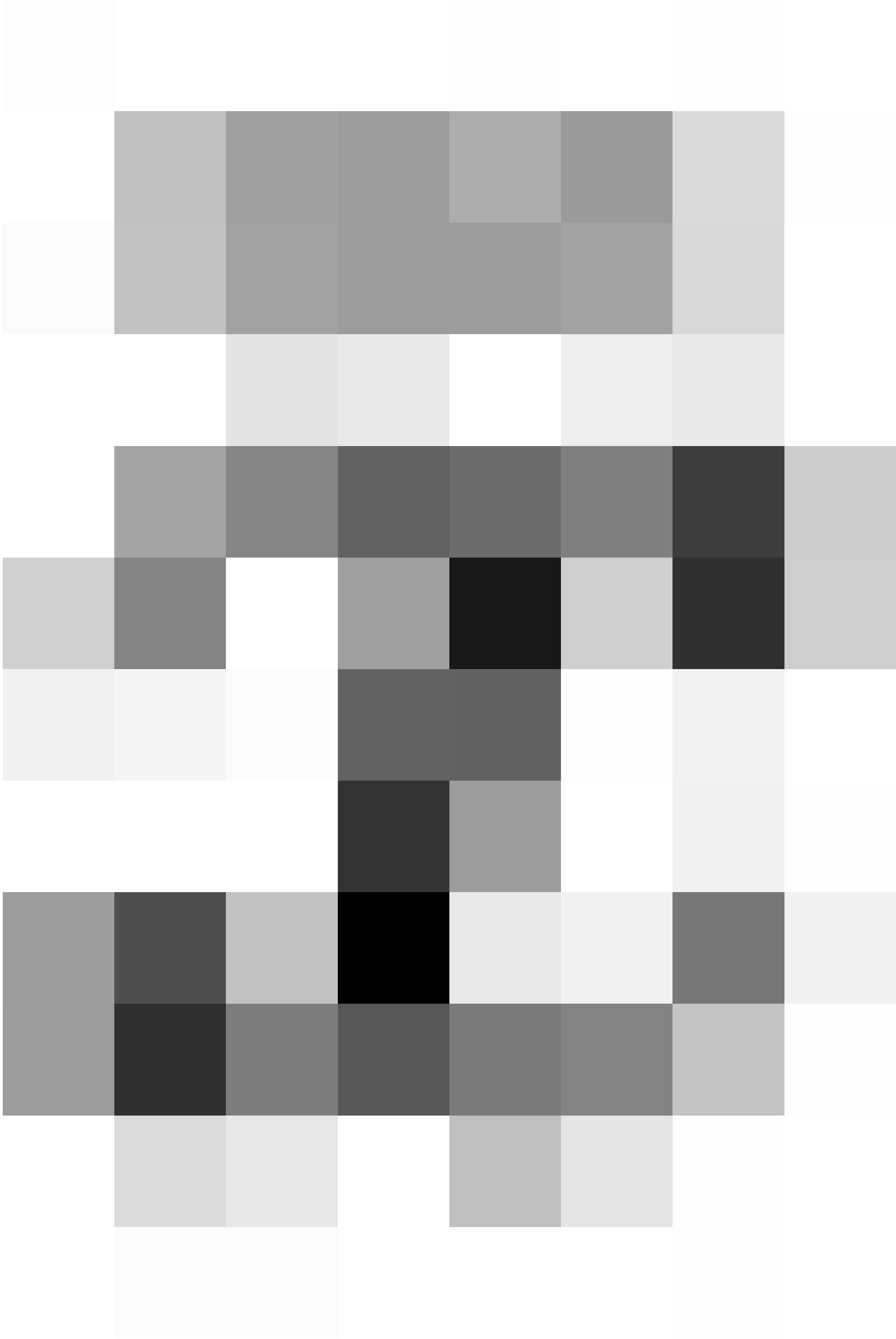


for samples of 25 observations.

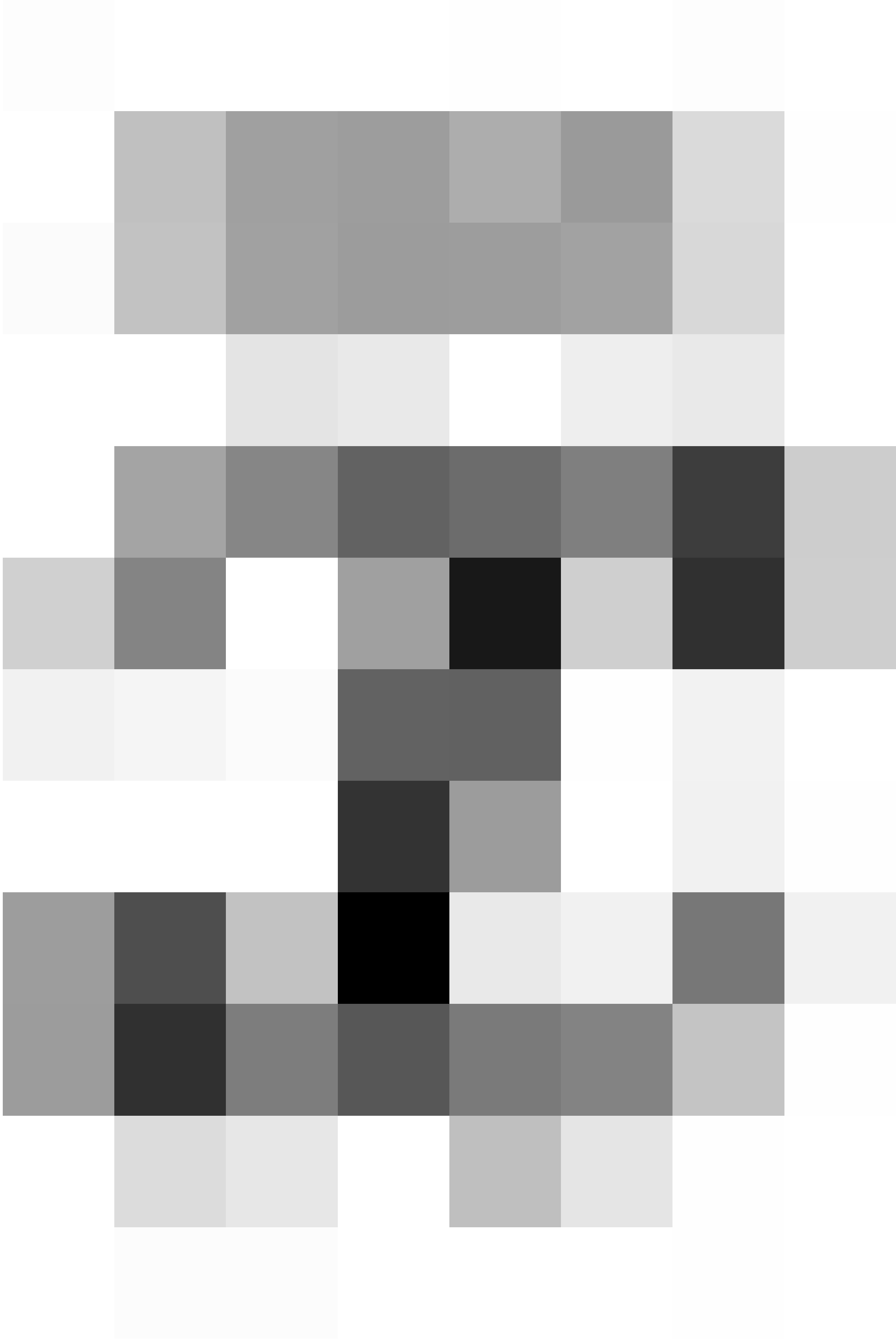
FIGURE 13.6 **(right)** The sampling distribution of $\bar{x}$



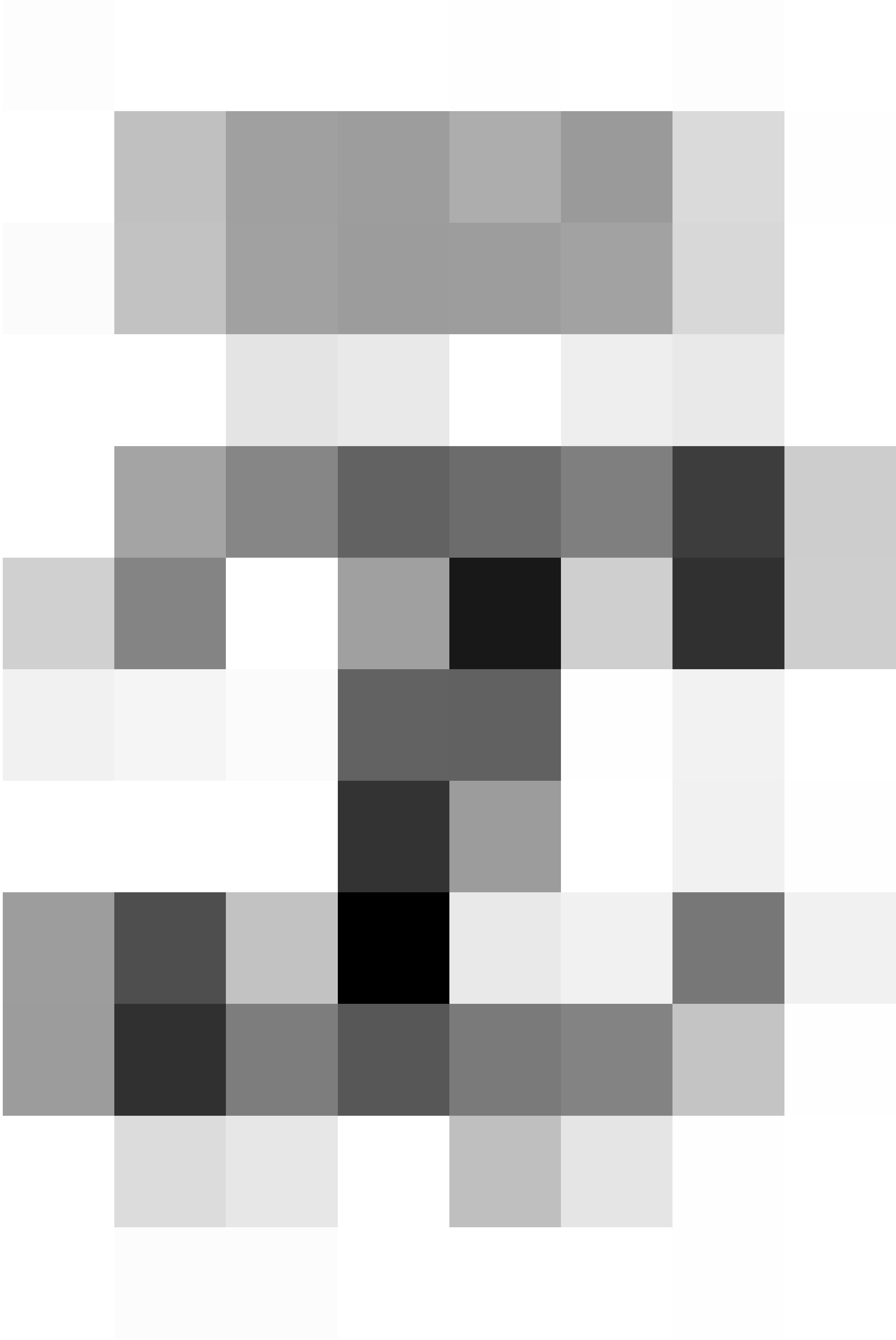
is always Normal when the population is Normal. a The distribution of single observations (the population distribution IQ scores for a particular age group). b The sampling distribution of $\bar{x}$



for samples of 2 observations. c The sampling distribution of $\bar{x}$



for samples of 10 observations. d The sampling distribution of $\bar{x}$



for samples of 25 observations.

But what about the *shape of a sampling distribution*? We saw in Figure 13.4 that the shape of the sampling distribution is Normal when the population distribution is Normal. This is also evident in the sampling distributions shown in Figure 13.6. But what if the population distribution is not Normal? It is a remarkable fact that when we consider larger sample sizes, the distribution of $\bar{x}$ changes shape if the population is non-Normal: It looks less like the population and more like a Normal distribution. When the sample size n is large enough, the sampling distribution of $\bar{x}$ is very close to Normal. This is true no matter what shape the population distribution has.[5] This famous fact of probability theory, which is called the *central limit theorem*, is extremely useful because of its wide applications.

CENTRAL LIMIT THEOREM

Draw an SRS of size n from any population with mean μ and standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal.

We can combine the central limit theorem with what we already know about the mean and standard deviation of the sampling distribution of $\bar{x}$. Therefore, when n is large enough, the sampling distribution of $\bar{x}$ follows approximately the $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ distribution. The central limit theorem allows us to use Normal probability calculations to answer questions about sample means from many observations (questions relying on the sampling distribution of the sample mean) even when the population distribution is not Normal.

EXAMPLE 13.6 The central limit theorem in action

Let's examine the shape of sampling distributions from the strongly skewed, definitely not Normal, exponential distribution of Figure 13.5(a). Figures 13.5(b), (c), and (d) show that the shape of the sampling distribution of $\bar{x}$ is closer to Normal for the larger values of n. The density curve for samples of 10 observations is still somewhat skewed to the right but already resembles a Normal curve having $\mu = 1$ and σ=1/10=0.32.$\sigma = 1/\sqrt{10} = 0.32.$ The density curve of the distribution of x¯$\bar{x}$ when n=25$n = 25$ is yet more Normal. The contrast between the shape of the population distribution and the shapes of the sampling distributions of the mean of x¯$\bar{x}$ for samples of 10 or 25 observations is striking.

The *Central Limit Theorem* applet on the companion website allows you to watch the central limit theorem in action, using computer simulations. Select the exponential distribution on the right and choose a sample size n.$n.$ The applet will simulate taking 10,000 random samples of size nn and plot the resulting 10,000 sample means. If you select sample sizes of 1 (individual observations from the population), 2, 10, and 25, you should obtain distributions with patterns similar to those shown in Figure 13.5.

APPLET

The central limit theorem applies to sampling distributions, not to the distribution (histogram) of the data from one random sample of n observations. Many students mistakenly believe that larger sample sizes yield more Normal sample histograms. This is not the case. If a population is skewed, for instance, chances are that a histogram of a random sample from that population will be skewed as well. The central limit theorem describes only what happens to the distribution of the averages from repeated samples of a given size.

CAUTION

How large a sample size n is needed for x¯$\bar{x}$ *to be close to Normal depends on the population distribution. More observations are required if the shape of the population distribution is far from Normal.* The following points are always true, though.

CAUTION

THINKING ABOUT SAMPLE MEANS

Means of random samples are **less variable** than individual observations.

Means of random samples are **more Normal** than individual observations.

Mathematically proving the central limit theorem is beyond the scope of an introductory statistics textbook. Instead, we will focus on its practical uses.

EXAMPLE 13.7 Practical use of the central limit theorem

Figure 13.7 shows the histograms for three real data sets. By observing the overall shape of a histogram, we can guess whether the population from which the data were taken might be approximately Normal, somewhat Normal, or definitely not Normal. From this observation and the size of the sample shown we can deduce whether the corresponding sampling distribution of $\bar{x}$ might be approximately Normal.

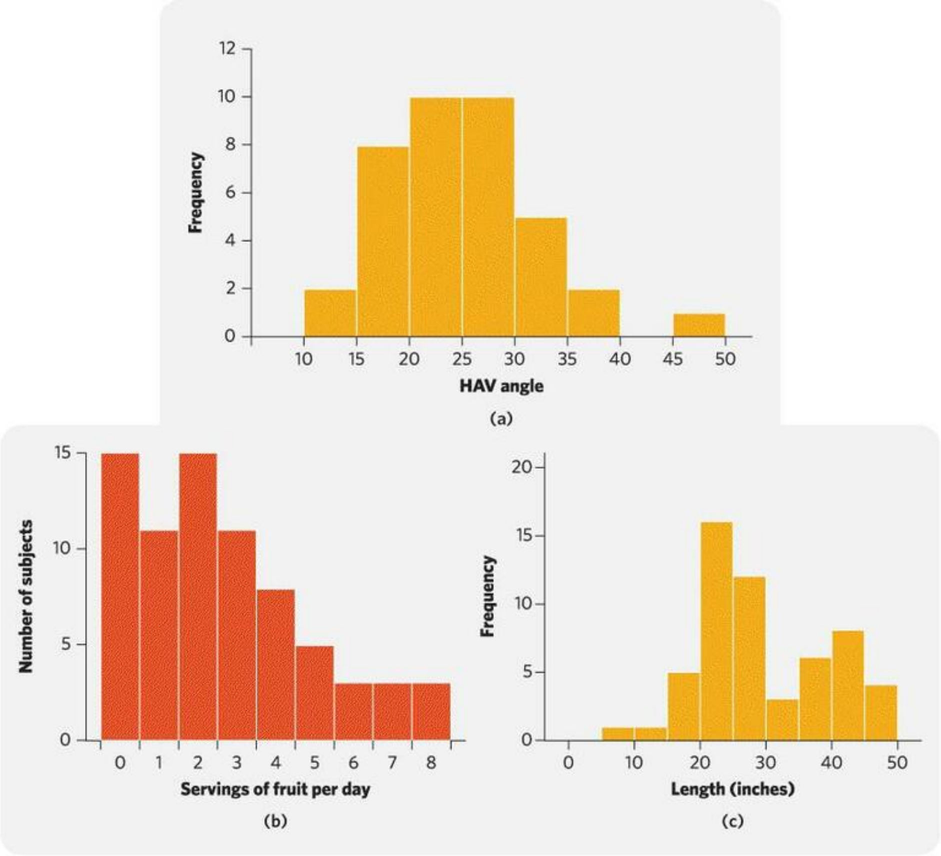


Figure 13.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 13.7 Histograms of three data sets for Example 13.6. (a) The angle of big-toe deformations (hallux abducto valgus [HAV]) in 38 patients. (b) The number of daily fruit servings for 74 adolescent girls. (c) The lengths of 56 perch from a Swedish lake.

Figure 13.7(a) shows the angle of big-toe deformations in 38 patients. The histogram is nearly Normal except for one mild outlier.

Because the outlier is not extreme and the sample size is reasonably large, the sampling distribution of $\bar{x}$ should be approximately Normal.

Figure 13.7(b) shows the number of daily fruit servings for 74 adolescent girls. The histogram is strongly right-skewed, suggesting that the population is not Normal but skewed. Even so, the sampling distribution of $\bar{x}$ should be approximately Normal, because the sample size is very large.

Figure 13.7(c) shows the lengths of 56 perch from a Swedish lake. The histogram is not particularly skewed and does not have any outliers. With a sample of 56 fish, the central limit theorem guarantees that the sampling distribution of $\bar{x}$ is approximately Normal. However, it is clear from the histogram that this sample is not homogeneous, but instead seems to be made up of two types of fish: some smaller fish with a peak at 20 to 25 inches, and some larger fish with a peak at 40 to 45 inches. The histogram does not reveal why this bimodal distribution might occur, and we would have to examine each fish closely to discover this reason. What is clear is that the sampling distribution would be centered on a value “μ” that would reflect neither of these two subpopulations of fish and, therefore, would be misleading. Keep in mind that poorly designed studies create problems that mathematics cannot remedy.

Sometimes it may be useful to look at published data to make an educated guess about the distribution of your own variable and decide on an appropriate sample size. For example, if you wanted to conduct your own study on the number of daily fruit servings among teenagers in your school district, you could rely on a publication showing the histogram of Figure 13.7(b) to guess that your own data would probably be skewed, too; this knowledge would help you decide to include at least 30 or 40 teenagers in your own sample to take advantage of the central limit theorem. Because sampling distributions of $\bar{x}$ become more Normal for larger sample sizes n, you have a little room to decide just how large a sample size you really need, depending on how accurate your Normal computations must be.

More general versions of the central limit theorem state that the distribution of any sum or average of many small random quantities is close to Normal. This is true even if the quantities are correlated with each other (as long as they are not too highly correlated) and even if they have different distributions (as long as no one random quantity is so large that it dominates the others). The central limit theorem suggests why Normal distributions are such common models for observed data. Any variable that is a sum of many small influences (for example, heights or IQ scores in a homogeneous population) will have an approximately Normal distribution.

APPLY YOUR KNOWLEDGE

13.8 What does the central limit theorem say? When asked what the central limit theorem says, a student replies, "As you take larger and larger samples from a population, the histogram of the sample values looks more and more Normal." Is the student right? Explain your answer.

13.9 Coffee consumption. A 2015 Gallup survey of a random sample of 1009 U.S. adults reports that "26% of American adults [in the sample] say they drink one cup of coffee on an average day, 19% drink two, 8% drink three, and 11% drink four or more. The remaining 36% drink none."

a. Describe the distribution of daily coffee consumption in the sample. (Hint: You can display the reported findings in a histogram.) What does it suggest about the shape of the distribution of daily coffee consumption in the entire U.S. adult population?

b. What can we say about the shape of the sampling distribution of average daily coffee consumption in random samples of 1009 U.S. adults? Explain your answer.

c. Researchers plan to recruit a random sample of 15 U.S. adults for a study of the impact of long-term use of vitamin supplements. Coffee consumption is a possible confounding variable. What can we say about the shape of the sampling distribution of average daily coffee consumption in random samples of 15 U.S. adults? Explain your answer.

pinkcandy/Depositphotos

13.10 Insurance. An insurance company knows that, in the entire population of millions of insured workers, the mean annual cost of worker’s compensation claims is $\mu=\$439$ $\mu = \$439$ per insured worker, and the standard deviation is $\sigma=\$20,000.$ $\sigma = \$20,000.$ The distribution of losses is strongly right-skewed: Most policies have no loss, but a few have large losses, up to millions of dollars.

a. If the company sells 40,000 policies, what are the shape, mean, and standard deviation of the sampling distribution of the mean claim loss? Consider these 40,000 policies to be a random sample of all worker’s compensation insurance policies.

b. If the company sells 40,000 policies, what is the probability that the mean claim loss would be no greater than \$500? No greater than \$700?

THE SAMPLING DISTRIBUTION OF $\hat{p}$

As illustrated in Examples 13.1 and 13.2, the sample average $\bar{x}$ is a statistic commonly used when studying quantitative data, and the sample proportion $\hat{p}$ is a statistic commonly used when dealing with categorical data. This section describes the sampling distribution of the statistic $\hat{p}$.

Suppose an experiment finds that 6 of 20 birds exposed to an avian flu strain develop flu symptoms. The number of birds that develop flu symptoms is a random variable *X*. *X* is a **count** of the occurrences of some categorical outcome in a fixed number of observations. If the number of observations is n, then the **sample proportion** is

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

EXAMPLE 13.8 Contact lens practices

Failure to follow recommended contact lens wear and care can lead to serious eye infection. A survey of a random sample of 281 Americans who wear contact lenses regularly asked about contact lens practices. The survey found that 139 respondents do not consistently wash hands before handling their contact lenses.

Comstock/Getty Images

The value 139 is the count *X* of regular contact lens wearers in the sample who do not consistently wash hands before handling their contact lenses. The proportion who do not consistently wash hands before handling their contact lenses among the 281 regular contact lens wearers in the sample is the sample proportion p^=139/281=49.5%. $\hat{p} = 139/281 = 49.5\%.$

The statistic 49.5% is an *estimate* of the population proportion pp of all American regular contact lens wearers who do not consistently wash hands before handling their contact lenses.

Categorical variables can take any of a finite number of possible outcomes. We may choose to call one such possible outcome a “success” and define all other possible outcomes as nonsuccesses, or failures. (This is an arbitrary decision, not a moral judgment.) If you have covered Chapter 12, you know that the distribution of the count *X* of a random “success” event occurring within nn observations is binomial when the probability pp of that event is constant over all observations and successive observations are independent. You also learned that the binomial distribution of *X* can be approximated by the Normal distribution N(np,np(1-p))$N\left(np, \sqrt{np(1-p)}\right)$ when nn is so large that the count of successes and the complementary count of failures in the sample are each at least 10.

A count of successes in itself is meaningful only in the context of the total number of observations, so it is of limited use when comparing different studies. That’s why we typically favor a more informative statistic, the sample proportion p^$\hat{p}$ of successes. How good is the statistic p^$\hat{p}$ as an estimate of the parameter p,$p,$ the proportion of successes in the population? To find out, we ask, “What would happen if we took many samples?”

EXAMPLE 13.9 What would happen in many samples?

The American Red Cross reports that 45% of Americans have blood type O. We plan to take a random sample of 1000 medical records. We can simulate what would happen if we took many such random samples. Figure 13.8 illustrates this process, from the population at the left, to choosing an SRS and finding the proportion p^$\hat{p}$ for this sample, to collecting and plotting the p^$\hat{p}$ values from many samples at the right.

Using Minitab’s “Random data” tool, we can find the count *X* of successes in a simulated SRS of 1000 observations from a population with probability p=0.45$p = 0.45$ of success. The corresponding sample proportion is p^=X/1000.$\hat{p} = X/1000.$ We can repeat this simulation many times. The first sample has p^=0.452.$\hat{p} = 0.452.$ The second sample contains a different set of 1000 observations, with p^=0.436,$\hat{p} = 0.436,$ and so on. The histogram at the right of Figure 13.8 shows the distribution of the values of p^$\hat{p}$ from 500 separate SRSs of size 1000. This histogram displays the underlying structure of the *sampling distribution* of the statistic p^.$\hat{p}.$

The sampling distribution of p^$\hat{p}$ is the pattern that would emerge if we looked at all possible random samples of size n=1000$n = 1000$ from our population. A simulation of just 500 SRSs, like that depicted in Figure 13.8, can give only an approximation of this sampling distribution. The actual properties of the sampling distribution of p^$\hat{p}$ can be derived mathematically, though the mathematical proofs are beyond the scope of an introductory statistics textbook. Here are the facts.

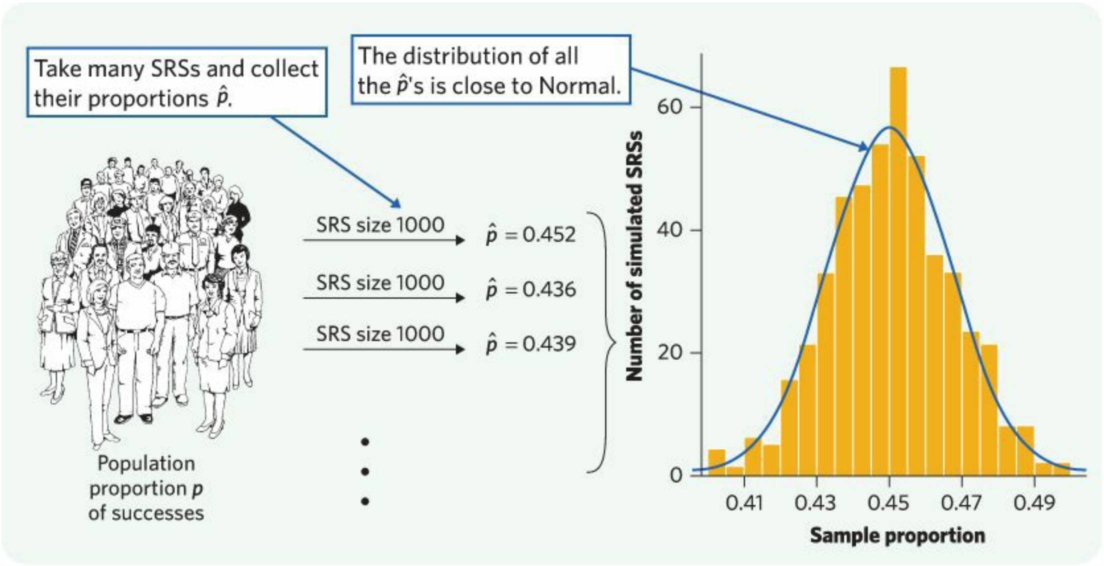
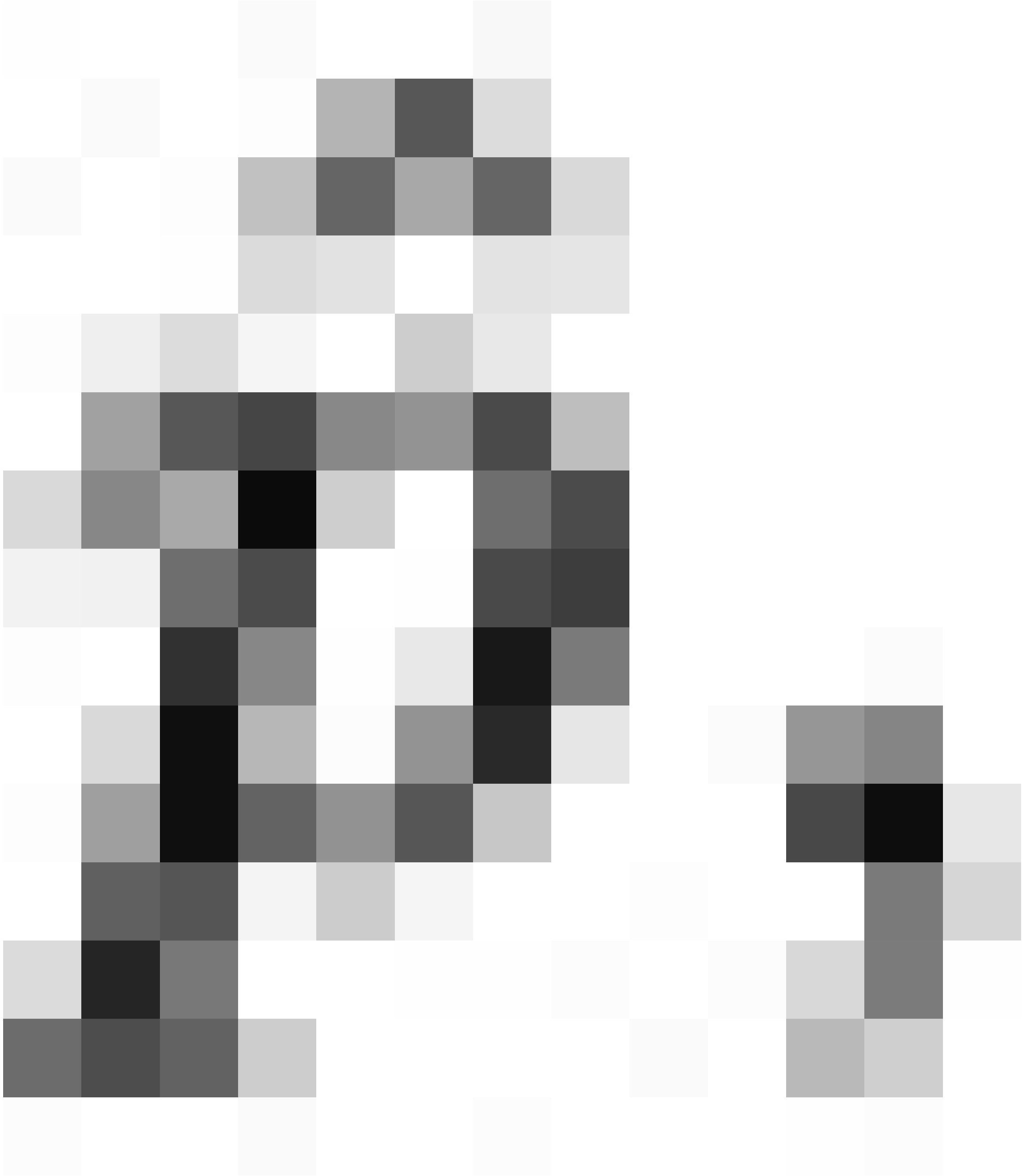
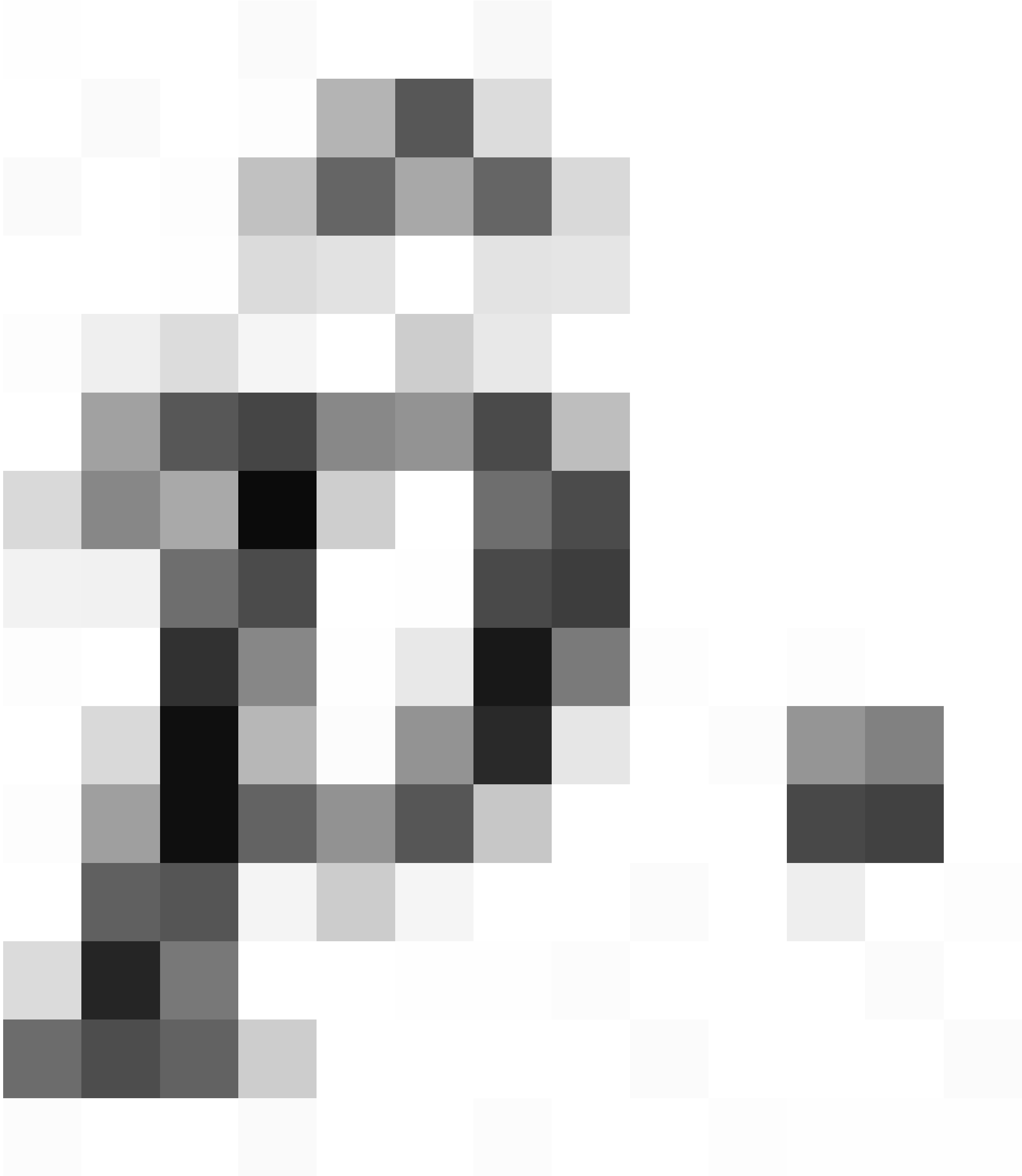


Figure 13.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 13.8 The idea of a sampling distribution: Take many SRSs from the same population, compute their sample proportions p^,



and display the distribution of p^.



The histogram shows the result of 500 simulated SRSs from the population of blood types, with proportion 0.45 having type O.

SAMPLING DISTRIBUTION OF A SAMPLE PROPORTION

Choose an SRS of size n from a large population that contains population proportion p of successes. Let $\hat{p}$ be the **sample proportion** of successes:

$$\hat{p} = \frac{\text{count of successes in the sample}}{n}$$

$$\hat{p} = \frac{\text{count of successes in the sample}}{n}$$

Then:

- The **mean** of the sampling distribution is p.p.
- The **standard deviation** of the sampling distribution is p(1-p)/n. $\sqrt{p\left(1-p\right)/n}$.
- As the sample size increases, the sampling distribution of p^$\hat{p}$ becomes **approximately Normal.**

Figure 13.9 displays the sampling distribution of p^.$\hat{p}$. The behavior of sample proportions p^$\hat{p}$ is similar to the behavior of sample means x¯.$\bar{x}$. When the sample size nn is large, the sampling distribution is approximately Normal. The larger the sample, the more nearly Normal the distribution is. The mean of the sampling distribution is the true value of the population proportion p.p. That is, p^$\hat{p}$ is an unbiased estimator of p.p. The standard deviation of p^$\hat{p}$ gets smaller as the sample size nn gets larger, so that estimation is likely to be more accurate when the sample is larger. As is the case for x¯,$\bar{x}$, the standard deviation gets smaller only at the rate n.$\sqrt{n}$. We need 100 times as many observations to cut the standard deviation by 10.

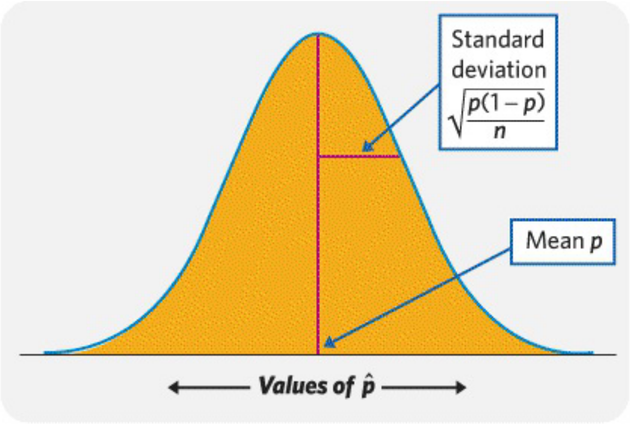
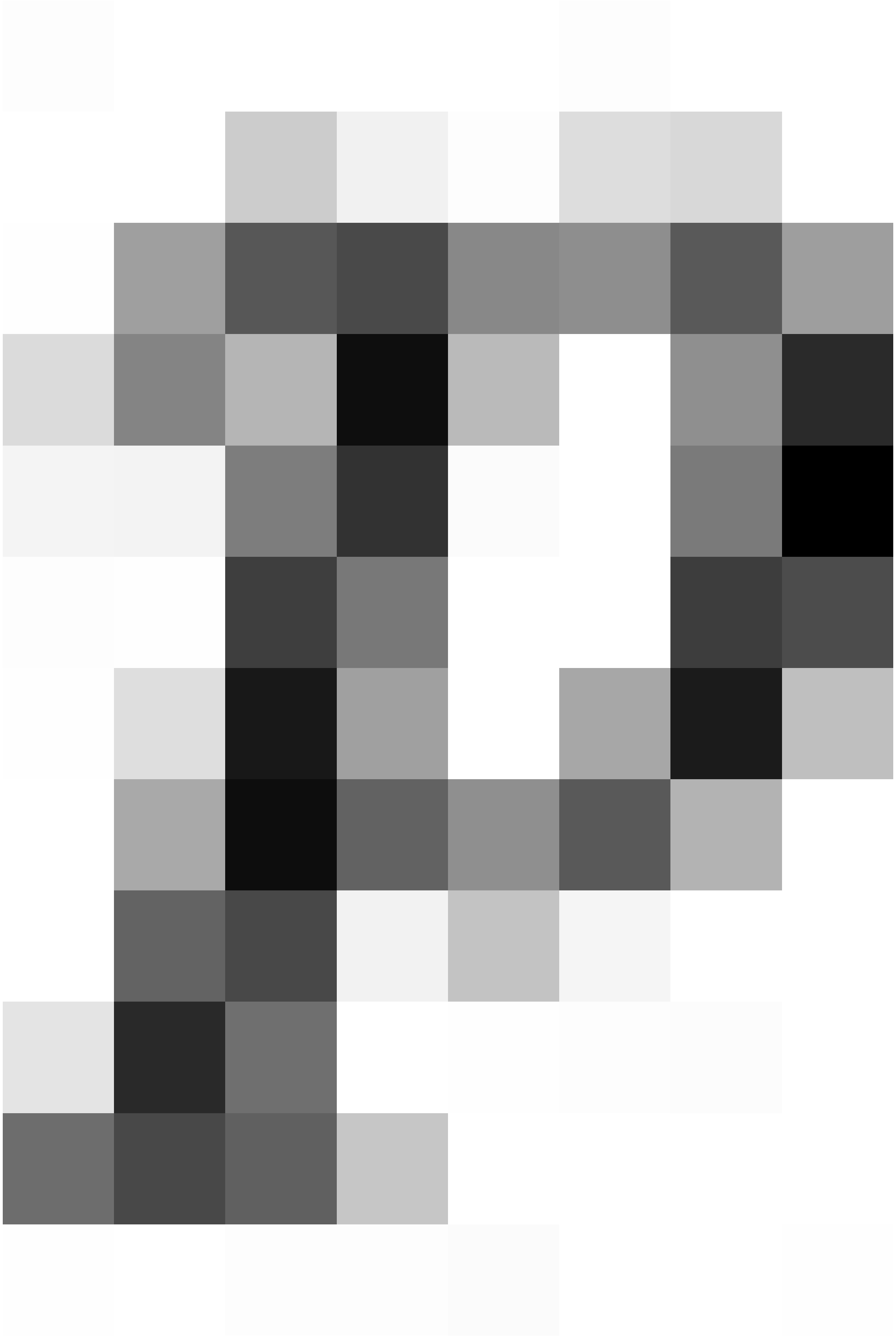
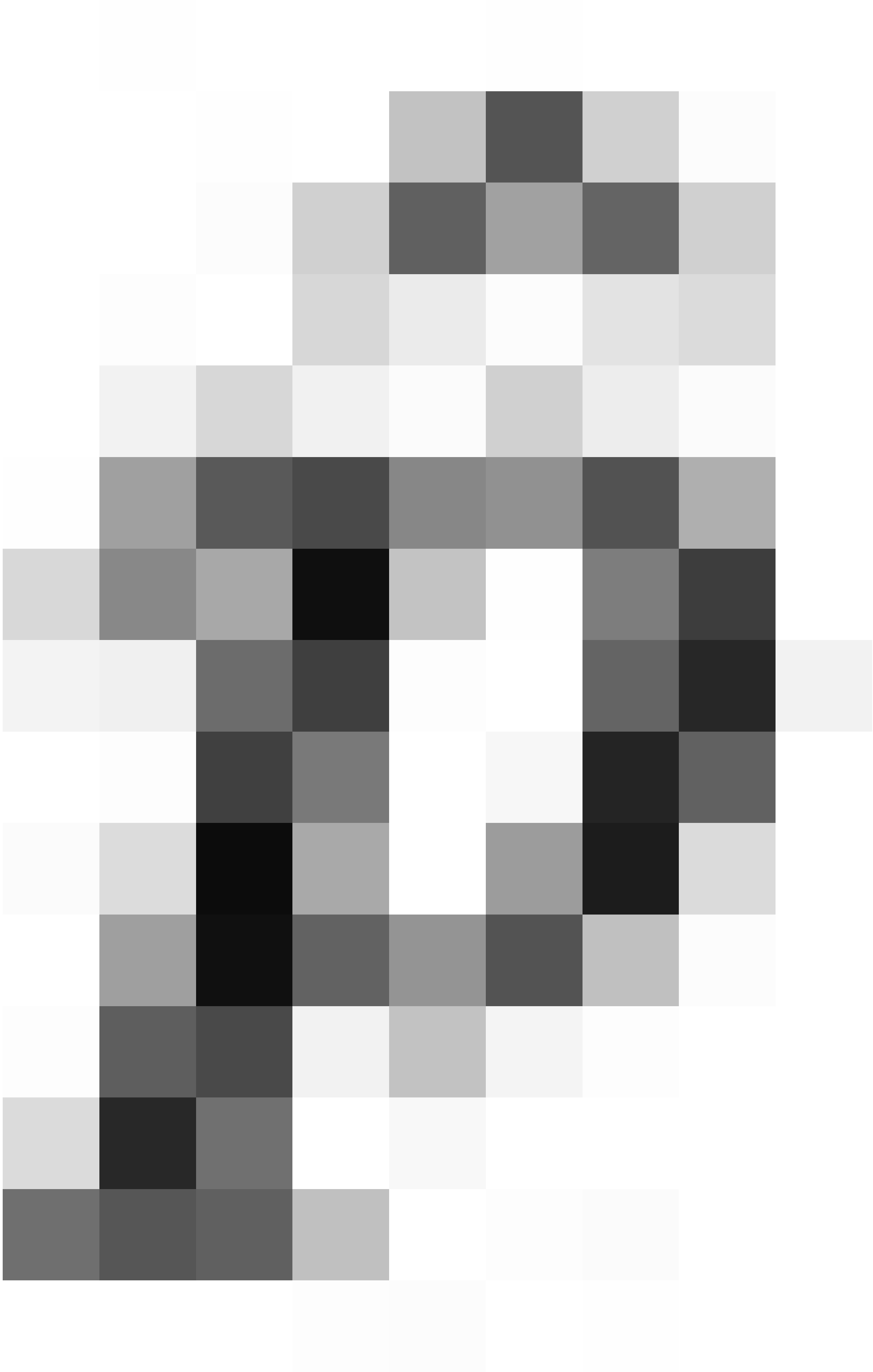


Figure 13.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 13.9 Select a large SRS from a population in which a proportion p



are successes. The sampling distribution of the proportion p^



of successes in the sample is approximately Normal. The mean is p,

$$p,$$

and the standard deviation is (p(1-p)/n).

$$\sqrt{(p(1-p)/n)}.$$

You should not use the Normal approximation for the distribution of p^ $\hat{p}$ *when the sample size* n n *is small. What is more, the formula given for the standard deviation of* p^ $\hat{p}$ *is not accurate unless the population is much larger than the sample—say, at least 20 times larger.*[6] We will give guidelines to help you decide when methods for inference based on this sampling distribution are trustworthy.

EXAMPLE 13.10 Blood types

The American Red Cross reports that 45% of Americans have blood type O. We plan to take a random sample of 1000 medical records. What is the probability that no more than 42% of the individuals in our sample would have blood type O?

Amanda Mills/CDC

If the sample size is n=1000$n = 1000$ and the population proportion is p=0.45,$p = 0.45,$ we know that the sampling distribution of the sample proportion p^$\hat{p}$ is approximately Normal, has mean 0.45, and has standard deviation

p(1-p)n=(0.45)(0.55)1000=0.01573.

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.45)(0.55)}{1000}} = 0.01573.$$

We want the probability that p^$\hat{p}$ is 0.42 or less. Using the software R, the function pnorm(0.42, mean = 0.45, sd = 0.01573) gives a probability of 0.0282 for P(p^≤0.42).$P(\hat{p} \le 0.42)$. If we took many, many random samples of 1000 medical records, only 2.8% of all these SRSs would contain at most 42% of records from individuals with blood type O.

The Normal approximation for the sampling distribution of p^ $\hat{p}$ *is least accurate when p is close to 0 or 1.* You can see that if p=0,$p = 0,$ any sample must contain only failures. That is, p^=0 $\hat{p} = 0$ every time and there is no Normal distribution in sight. Similarly, the approximation works poorly when pp is close to 1. In practice, for the

CAUTION

Normal approximation, we need to select a larger sample size nn for values of pp near 0 or 1.

APPLY YOUR KNOWLEDGE

13.11 Aging American population. The 2010 U.S. census shows that **6%** of all Americans are at least 75 years of age. To test a random digit dialing device, you use the device to call randomly chosen residential telephones in your county. Of the 150 members of the households contacted, **4%** are 75 years or older.

a. Is each of the boldface numbers a parameter or a statistic?
b. What are the mean and standard deviation of the proportion p^$\hat{p}$ who are at least 75 years of age in samples of 150 U.S. respondents?

13.12 Baby boys. We saw in Example 9.1 (page 216) that, historically, approximately 51% of the babies born in the United States are boys. That is, we know that p=0.51.$p = 0.51.$ A survey plans to study 650 randomly selected newborns.

a. What are the mean and standard deviation of the statistic p^,$\hat{p},$ the sample proportion who are boys?
b. Use the Normal approximation to find the probability that p^$\hat{p}$ would be less than 0.50.

13.13 Aging American population, continued. Based on the information presented in Exercise 13.11, what is the probability that a random sample of 150 respondents would contain less than 4% of individuals at least 75 years of age?

13.14 Baby boys, continued. Exercise 13.12 asks you to find the probability that the sample proportion p^$\hat{p}$ is less than 0.5 in a sample of 650 newborns. Find this probability for SRSs of sizes 1000, 4000, and 16,000. What general fact do your results illustrate?

THE LAW OF LARGE NUMBERS[i]

Looking at the properties of the sampling distributions for x¯ $\bar{x}$ and for p^, $\hat{p}$, we notice that their standard deviations are a function of the chosen sample size n:n: σ/n$\sigma/\sqrt{n}$ and p(1-p)/n,$\sqrt{p(1-p)/n}$, respectively. In particular, we can see that for a very large sample size n,n, the standard deviation of a sampling distribution must be very small. This means that if we keep taking larger and larger samples, the statistic x¯ $\bar{x}$ is *guaranteed* to get closer and closer to the parameter μμ and the statistic p^$\hat{p}$ is *guaranteed* to get closer and closer to the parameter p.p. This remarkable fact is called the **law of large numbers.**

APPLET

Figure 9.1 (page 217) shows how proportions approach a given probability in a coin-tossing example. The *Law of Large Numbers* applet helps illustrate this concept for a quantitative variable using dice. You can use the applet to watch x¯ $\bar{x}$ change as you average more and more observations until it eventually settles down at the mean μ.μ.

CAUTION

The law of large numbers is usually impractical for direct applications in the life sciences. Indeed, it is only *in the very long run* that the mean outcome is predictable, and studies in the life sciences rarely deal with such large numbers. However, the law of large numbers is the foundation of such businesses as gambling casinos and insurance companies. The winnings (or losses) of a gambler on a few plays are uncertain—that’s why, unfortunately, gambling can be addictive. The house plays tens of thousands of times. So the house, unlike individual gamblers, can count on the long-run regularity described by the law of large numbers. The average winnings of the house on tens of thousands of plays will be very close to the mean of the distribution of winnings, guaranteeing the house a profit. The same idea applies to health or property insurance. Most insured individuals will make no claim at all, yet some will have very expensive claims. An insurance company insuring hundreds of thousands of customers can count on the law of large numbers to know that the average amount it will pay per customer in a given year will be very close to the expected mean claim price. For other scenarios, however, the law of large numbers is impractical.

APPLY YOUR KNOWLEDGE

13.15 More on insurance. The idea of insurance is that we all face risks that are unlikely to be realized but carry high costs. Worker's compensation insurance provides for the cost of medical care, rehabilitation, and other related expenses for injured workers. In a given year, most workers will not have any work injury, some will have minor injuries, and a few will die or become permanently disabled. Insurance spreads the risk: We all pay a small amount, and the insurance policy pays a large amount to these few extreme claims. The National Academy of Social Insurance reports that the mean worker's compensation claim cost in a year is $\mu=\$439$$\mu = \$439$ per covered worker. An insurance company plans to sell worker's compensation insurance for $439 plus enough to cover its administrative costs and profit. Explain clearly why it would be unwise to sell only 12 policies. Then explain why selling thousands of such policies is a safe business.

13.16 Means in action. Use the *Law of Large Numbers* applet. The applet simulates successive rolls of a die and recomputes the average number of spots on the up-face after each roll. Over the long run, the mean number of spots is 3.5.

APPLET

a. Type 20 in the "Rolls" window. Then click on the "Roll dice" button. What do you observe? Reset the applet and repeat this simulation three more times. How are these four simulations different? What do they have in common? You can expect a lot of variability for such small sample sizes.

b. Now type 100 in the "Rolls" window and click on the "Roll dice" button three times so that you have a total of 300 rolls. What do you observe? Reset the applet and repeat this simulation one more time. How are these two simulations different? What do they have in common? Although 300 is not a huge number, you can still get an idea of how the law of large numbers works.

CHAPTER 13 SUMMARY

- A **parameter** is a number that describes a population, whereas a **statistic** is a number that describes a sample.
- When we want information about the **population mean** μμ for some quantitative variable, we often take an SRS and use the **sample mean** x‾ $\bar{x}$ to estimate the unknown parameter μ.μ.
- The **sampling distribution** of x‾ $\bar{x}$ describes how the statistic x‾ $\bar{x}$ varies in all possible SRSs of the same size from the same population.
- The **mean** of the sampling distribution of x‾ $\bar{x}$ is the same as the population mean μ,μ, so x‾ $\bar{x}$ is an **unbiased estimator** of μ.μ.
- The **standard deviation** of the sampling distribution of x‾ $\bar{x}$ is equal to σ/n$\sigma/\sqrt{n}$ for an SRS of size nn if the population has standard deviation σ.σ. That is, averages are less variable than individual observations.
- If the population has a Normal distribution, so does the sampling distribution of x‾ .$\bar{x}$.
- The **central limit theorem** states that for large nn the sampling distribution of x‾ $\bar{x}$ is approximately Normal regardless of the shape of the population distribution. That is, averages are more Normal than individual observations. We can use the N(μ,σ/n)$N(\mu, \sigma/\sqrt{n})$ distribution to calculate approximate probabilities for events involving x‾ .$\bar{x}$.
- When we want information about the **population proportion** pp for some categorical variable, we often take an SRS and use the **sample proportion** p^$\hat{p}$ to estimate the unknown parameter p.p.
- The **sampling distribution of the proportion** p^$\hat{p}$ **of successes** in an SRS of size nn from a large population containing proportion pp of successes has a mean equal to pp and a standard deviation equal to p(1-p)/n.$\sqrt{p(1-p)/n}$.
- The **mean** of the sampling distribution of p^$\hat{p}$ is p,p, and therefore p^$\hat{p}$ is an **unbiased estimator** of p.p.

- When the sample size nn is large, when both npnp and n(1-p)$n(1-p)$ are large enough, and when the population is at least 20 times larger than the sample, the sampling distribution of p^$\hat{p}$ is approximately Normal. We can use the N(p,p(1-p)/n)$N\left(p, \sqrt{p(1-p)/n}\right)$ distribution to calculate approximate probabilities for events involving p^.$\hat{p}$.
- The **law of large numbers** states that the observed mean outcome x¯ $\bar{x}$ must approach the mean μμ of the population as the number of observations increases. Likewise, the observed proportion p^$\hat{p}$ of a categorical outcome must approach the proportion of this outcome in the population as the number of observations increases.

THIS CHAPTER IN CONTEXT

This chapter brings together a number of concepts from previous chapters. We have examined random variables, their probability distribution, their mean, and their standard deviation. We now apply these same concepts to the outcome of random samples. We also consider two distinct situations that depend on the type of data collected, quantitative or categorical.

Sampling distributions are closely related conceptually to random samples and populations. It is therefore essential to understand the distinctions between these three concepts. We first examine the case of a quantitative variable, x.

Population. A population contains all the individuals making up a group of interest, along with their associated numerical value for the variable x. We rarely have information about each individual in the population (there are too many to examine). Instead, we use a density curve to model the probability distribution of the variable x studied in the population, as we saw in Chapters 9 and 11. Normal curves, described in Chapter 11, are one type of density curve that we can sometimes use to model a population distribution. The population mean is labeled μ, and the population standard deviation is labeled μ. (Sometimes we will be given the values of μ and σ, but in most cases we will not know them.)

Random sample. A random sample is a subset of n individuals from a population of interest, along with their associated numerical value x. The sample can be obtained by randomly selecting individuals from the population or by randomly assigning individuals to experimental conditions, as seen in Chapters 7 and 8. We do have data for these n individuals, and we can plot the data in a histogram or a dotplot (Chapter 1). In Chapter 2, we saw how to compute summary statistics, such as the mean $\bar{x}$ and the standard deviation *s* of sample data.

Sampling distribution. The sampling distribution of $\bar{x}$ contains all the means $\bar{x}$ for all possible random samples of size n from a population of interest. We do not usually have information about each sample mean (too many to examine). Instead, we use a density curve to model the probability distribution of the variable $\bar{x}$ for samples of size n

n from this population. The mean of the sampling distribution of x¯$\bar{x}$ is equal to the population mean μ,$\mu,$ and the standard deviation of this sampling distribution is equal to σ/n$\sigma/\sqrt{n}$ (the population standard deviation σσ divided by the square root of the sample size nn). We can use a Normal curve to model the sampling distribution of x¯$\bar{x}$ when the population is Normally distributed. Because of the central limit theorem, we can also use a Normal curve to model the sampling distribution of x¯$\bar{x}$ when the sample size nn is large enough. The central limit theorem does not apply to one random sample—it applies only to the shape of a sampling distribution.

When dealing with a categorical variable, we can define one particular outcome of interest as a “success” and any other outcome as a “failure.” This approach should be familiarif you have read optional Chapter 12. The population contains all the individuals making up a group of interest, along with their associated outcome (success or failure). The population parameter is the proportion of successes pp in the population. (The value of this parameter is sometimes given, but most often it is unknown.) For a random sample, we can compute the sample proportion p^$\hat{p}$ of successes. The sampling distribution of p^$\hat{p}$ has a mean equal to pp and a standard deviation equal to p(1-p)/n.$\sqrt{p(1-p)/n}.$ We can use a Normal distribution to model the sampling distribution of p^$\hat{p}$ when the sample size nn is large enough.

Sampling distributions may seem like strange theoretical ideas, but they have tremendous applications. In the next two chapters, we will examine how the sampling distribution of x¯$\bar{x}$ allows us to perform statistical inference about the population parameter μ.$\mu.$ Later, in Chapter 19, we will see how the sampling distribution of p^$\hat{p}$ lets us perform statistical inference about the population parameter p.$p.$

CHECK YOUR SKILLS

13.17 The National Center for Health Statistics interviewed all adult household members of a nationally representative sample of the civilian noninstitutionalized household population; **62%** of the interviewees had used some form of complementary or alternative medicine, including prayer, during the past 12 months. The boldface number is a

a. sampling distribution.
b. parameter.
c. statistic.

13.18 A newborn baby has extremely low birth weight (ELBW) if it weighs less than 1000 grams. A study of the health of such children in later years examined a random sample of 219 children. Their mean weight at birth was $\bar{x} = 810$ grams. This sample mean is an *unbiased estimator* of the mean weight μ in the population of all ELBW babies. This means that

a. in many samples from this population, the mean of the many values of $\bar{x}$ will be equal to μ.
b. as we take larger and larger samples from this population, $\bar{x}$ will get closer and closer to μ.
c. in many samples from this population, the many values of $\bar{x}$ will have a distribution that is close to Normal.

Cholesterol levels among fourteen-year-old boys are roughly Normal, with mean 170 and standard deviation 30 milligrams per deciliter (mg/dl). Use this information for Exercises 13.19 to 13.21.

13.19 You choose an SRS of 4 fourteen-year-old boys and average their cholesterol levels. If you do this many times, the mean of all the average cholesterol levels you get will be close to

a. 170 mg/dl.
b. $170/4 = 42.5$ mg/dl.
c. $170/\sqrt{4} = 85$ mg/dl.

13.20 You choose an SRS of 4 fourteen-year-old boys and average their cholesterol levels. If you do this many times, the standard deviation of all the average cholesterol levels you get will be close to

a. 30 mg/dl.$30\ \mathrm{mg/dl}.$
b. 4/30=0.73 mg/dl.$4/\sqrt{30} = 0.73\ \mathrm{mg/dl}.$
c. 30/4=15 mg/dl.$30/\sqrt{4} = 15\ \mathrm{mg/dl}.$

13.21 In an SRS of 4 fourteen-year-old boys, the probability that the average cholesterol level is 200 mg/dl or more is close to

a. 0.023.
b. 0.159.
c. 0.977.

13.22 The survival times of guinea pigs inoculated with an infectious viral strain vary from animal to animal. The distribution of survival times is strongly skewed to the right. The central limit theorem says that

a. as we study more and more infected guinea pigs, their average survival time gets ever closer to the mean μμ for all infected guinea pigs.
b. the average survival time of a large number of infected guinea pigs has a distribution of the same shape (strongly skewed) as the distribution for individual infected guinea pigs.
c. the average survival time of a large number of infected guinea pigs has a distribution that is close to Normal.

According to the CDC, the proportion of all gonorrhea cases tested in 2014 that were resistant to a major antibiotic was 0.38.[7] *A researcher plans to take an SRS of 100 medical records of patients diagnosed with gonorrhea in 2014 and obtain the proportion that showed resistance to a major antibiotic. Use this information for Exercises 13.23 to 13.25.*

13.23 The number 0.38 is a

a. sampling distribution.
b. parameter.
c. statistic.

13.24 If you took many different samples of 100 medical records with a gonorrhea diagnosis in 2014, the standard deviation of the proportion of gonorrhea cases that were resistant to a major antibiotic would be close to

a. 0.38$0.38$
b. 0.38×0.62/100=0.0236.$0.38 \times 0.62/\sqrt{100} = 0.0236.$
c. 0.38×0.62/100=0.0485.$\sqrt{0.38 \times 0.62/100} = 0.0485.$

13.25 In an SRS of 100 medical gonorrhea cases from 2014, the probability of getting 40% or more cases that are resistant to a major antibiotic is close to

a. 0.198.
b. 0.340.
c. 0.660.

13.26 (Optional) Medical care and compensation costs for workers injured on the job vary a lot and are strongly skewed as a whole. The National Academy of Social Insurance reports that the mean worker's compensation claim cost in a year is μ=439\mu = \$439$ per insured worker, with standard deviation σ=$20,000.$\sigma = \$20,000.$ The law of large numbers says that

a. an insurance company can get an average worker's compensation claim cost that is lower than the mean $439 by insuring a large number of workers.
b. as a company insures more and more workers chosen at random, the average claim cost gets ever closer to $439.
c. if a company insures a large number of workers chosen at random, that company's average claim cost will have approximately a Normal distribution.

CHAPTER 13 EXERCISES

13.27 Women’s heights. A random sample of female college students has a mean height of **65** inches, which is greater than the **64**-inch mean height of all young women. Is each of the bold numbers a parameter or a statistic? Explain your answer.

13.28 Lightning strikes. The number of lightning strikes on a square kilometer of open ground in a year has mean 6 and standard deviation 2.4. (These values are typical of much of the United States.) The National Lightning Detection Network uses automatic sensors to watch for lightning in a sample of 25 randomly chosen square kilometers. What are the mean and standard deviation of x¯ ,$\bar{x}$, the mean number of strikes per square kilometer?

13.29 Heights of male students. Suppose that the distribution of heights of all male students on your campus is Normal, with mean 70 inches and standard deviation 2.8 inches.

a. If you choose one student at random, what is the probability that he is between 69 and 71 inches tall?
b. What is the standard deviation of x¯ ,$\bar{x}$, the mean height for a sample of nn male students? How large an SRS must you take to reduce the standard deviation of the sample mean to 0.5 inch?
c. What is the probability that the mean height of the sample in part b is between 69 and 71 inches?

13.30 Lightning strikes, continued. The number of lightning strikes on a square kilometer of open ground in a year has mean 6 and standard deviation 2.4. What is the probability that the average number of lightning strikes per square kilometer per year for 25 randomly chosen square kilometers is 4.8 or less?

13.31 Heights of male students, continued. Suppose that the distribution of heights of all male students on your campus is Normal, with mean 70 inches and standard deviation 2.8 inches.

a. What standard deviation must x¯ $\bar{x}$ have so that 99.7% of all samples give an x¯ $\bar{x}$ within 0.5 inch of μ?μ? (Use the 68–95–99.7 rule.)
b. How large an SRS do you need to reduce the standard deviation of x¯ $\bar{x}$ to the value you found in part a?

13.32 Glucose testing. Shelia’s doctor is concerned that she may suffer from gestational diabetes (high blood glucose levels during pregnancy). There is variation both in the actual glucose level and in the blood test that measures the level. A patient is classified as having gestational diabetes if her glucose level is greater than 140 milligrams per deciliter (mg/dl) one hour after ingesting a sugary drink. Shelia’s measured glucose level one hour after the sugary drink varies according to the Normal distribution with μ=125 mg/dl$\mu = 125\,\mathrm{mg/dl}$ and σ=10 mg/dl.$\sigma = 10\,\mathrm{mg/dl}$.

a. If a single glucose measurement is made, what is the probability that Shelia is diagnosed with gestational diabetes?
b. If measurements are made on 4 different days and the mean result is compared with the criterion 140 mg/dl, what is the probability that Shelia is diagnosed with gestational diabetes?

13.33 Blood potassium level. Judy’s doctor is concerned that she may have hypokalemia (low potassium in the blood, measured in millimoles per liter, mmol/l). There is variation both in the actual potassium level and in the blood test that measures the level. Judy’s measured potassium level varies according to the Normal distribution with μ=3.8 mmol/l$\mu = 3.8\,\mathrm{mmol/l}$ and σ=0.2 mmol/l.$\sigma = 0.2\,\mathrm{mmol/l}$. A patient is classified as hypokalemic if the potassium level is less than 3.5 mmol/l.$3.5\,\mathrm{mmol/l}$.

a. If a single potassium measurement is made, what is the probability that Judy is diagnosed as hypokalemic?
b. If measurements are made instead on 4 separate days and the mean result is compared with the criterion of 3.5 mmol/l, what is the probability that Judy is diagnosed as hypokalemic?

13.34 Glucose testing, continued. Shelia’s measured glucose level one hour after ingesting a sugary drink varies according to the Normal distribution with $\mu=125$ mg/dl$\mu = 125\ \mathrm{mg/dl}$ and $\sigma=10$ mg/dl.$\sigma = 10\ \mathrm{mg/dl}.$ Let’s consider what could happen if we took 4 separate measurements from Shelia. What is the blood glucose level *L* such that the probability is only 0.05 that the average of 4 measurements is larger than *L*? (Hint: This requires an inverse Normal calculation.)

13.35 Blood potassium level, continued. Judy’s measured potassium level varies according to the Normal distribution with $\mu=3.8$$\mu = 3.8$ and $\sigma=0.2$ mmol/l.$\sigma = 0.2\ \mathrm{mmol/l}.$ Let’s consider what could happen if we took 4 separate measurements from Judy. What is the blood potassium level *L* such that the probability is only 0.05 that the average of 4 measurements is less than *L*? (Hint: This requires an inverse Normal calculation.)

13.36 Pregnancy lengths. The length of human pregnancies from conception to birth varies according to a distribution that is approximately Normal, with mean 266 days and standard deviation 16 days. A study enrolls a random sample of 16 pregnant women.

a. What are the mean and standard deviation of the sampling distribution of $\bar{x}$ $\bar{x}$?
b. What is the probability that the average pregnancy length exceeds 270 days?

13.37 Age at death. The World Health Organization reports that the distribution of ages at time of death in the United States has a mean $\mu$$\mu$ of 79.8 years and a standard deviation $\sigma$$\sigma$ of 15.5 years, based on death certificates. The distribution of ages at time of death is obviously left-skewed (thankfully, few people die young and most die old). Researchers plan to take a random sample of 100 American death certificates.

a. What are the mean and standard deviation of the sampling distribution of the mean age at death, $\bar{x}$,$\bar{x},$ for samples of 100 American death certificates?
b. Explain why the sampling distribution of $\bar{x}$ $\bar{x}$ is approximately Normal.
c. What is the probability that the mean age at time of death in a random sample of 100 American death certificates will be less than 78 years?

13.38 Blood alcohol content. The distribution of blood alcohol content in evening drivers is very skewed: Most drivers don’t drink and drive, some limit their drinking, and a few are intoxicated. If you wanted to make Normal calculations using the sampling distribution of the average blood alcohol content of evening drivers, would a random sample of 10 drivers be sufficient? 50 drivers? 200 drivers? Explain your answers.

13.39 Smoking and pregnancy. Cotinine, the metabolized form of nicotine, crosses the protective placental barrier and actually reaches the fetus. Traces can be found in a newborn’s first stool at birth, which is called meconium. You want to study cotinine levels in meconium, but you know from a review of the scientific literature that they are typically very strongly right-skewed. Your statistical analysis would benefit from an approximately Normal sampling distribution. What minimum sample size would be advisable for your study? Explain your reasoning.

13.40 Extrasensory perception. In tests for extrasensory perception (ESP), an experimenter looks at cards that are hidden from the subject. Each card contains one of five symbols, and the subject must name the symbol on each card presented. A subject with no suspected psychic ability has, by chance alone, a 20% chance of correctly guessing the symbol on each card presented. Because the cards are independent, the overall proportion of success can be expected to be **20%**. You have no psychic ability and score **28%** correct guesses out of a 25-card deck. Is each of the bold values a parameter or a statistic? Explain your answer.

13.41 Sampling bias. One way of checking for the effects of undercoverage, nonresponse, and other sources of error in a sample survey is to compare the sample with known demographic facts about the population. The 2010 census found that 13.9%, or 32,576,000, of the 235,016,000 adults (aged 18 and older) in the United States identified themselves as being of Hispanic origin. Is the value 13.9% a parameter or a statistic? Explain your answer.

13.42 STDs. To track epidemics, the CDC requires physicians to report all cases of important transmissible diseases. In 2014, for instance, a total of 350,062 cases of the sexually transmitted disease (STD) gonorrhea were officially reported, 53% of which were individuals in their 20s. Assume that the value 53% remains the same every year. Researchers plan to take a random sample of 400 diagnosed cases of gonorrhea to study the risk factors associated with the disease. Call p^$\hat{p}$ the proportion of cases in the sample corresponding to individuals in their 20s.

a. What are the mean and standard deviation of the sampling distribution of p^$\hat{p}$ in random samples of size 400?
b. Explain why, under these conditions, the sampling distribution of p^$\hat{p}$ is approximately Normal.

13.43 Sampling bias, continued. The 2010 census found that 13.9% of adults in the United States identified themselves as being of Hispanic origin.

a. An opinion poll plans to interview 1500 adults at random. What are the mean and standard deviation of the sampling distribution of the proportion of individuals of Hispanic origin for samples of this size?
b. Use a Normal approximation to find the probability that such a sample will contain 12% or fewer individuals of Hispanic origin.

13.44 STDs, continued. Based on the information presented in Exercise 13.42, use a Normal approximation to find the probability that no more than half of the 400 gonorrhea cases sampled would be from individuals in their 20s.

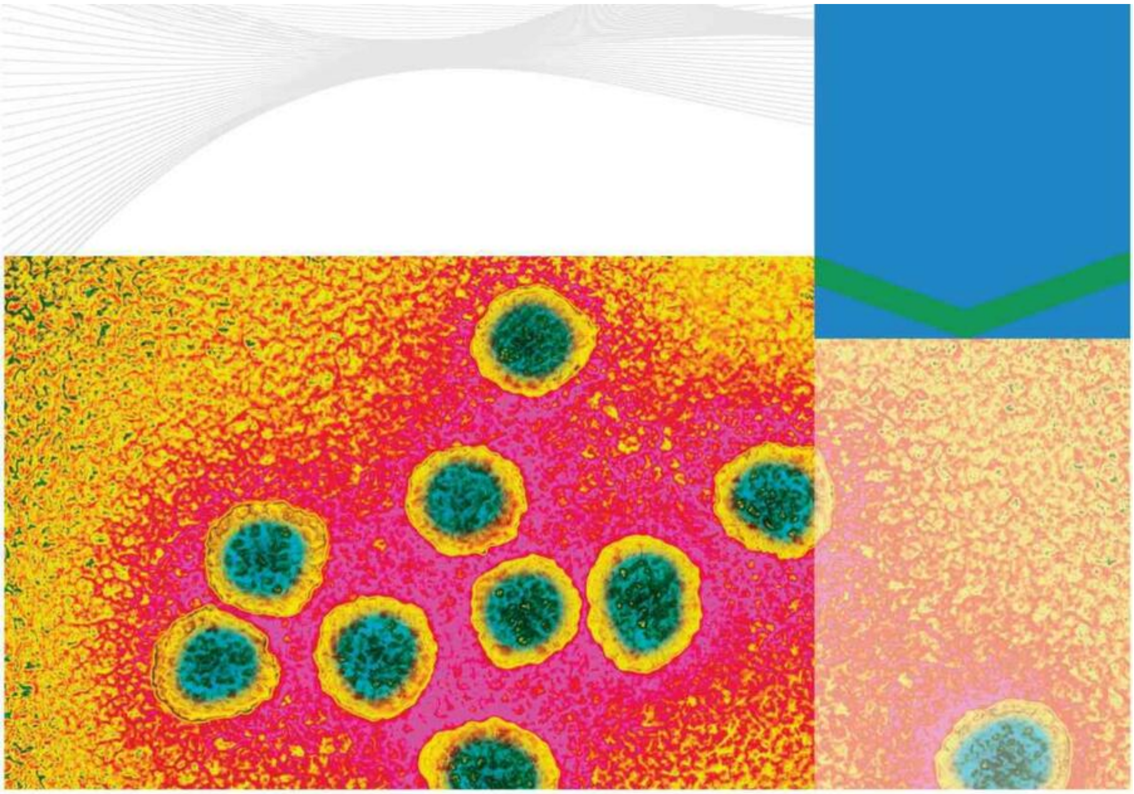
13.45 Contraception and unintended pregnancies. Pharmaceutical companies advertise that the birth control pill has an efficacy of 99.5% in preventing pregnancy. However, under typical use, the real efficacy is only approximately 95%. That is, 5% of women taking the pill for a year will experience an unplanned pregnancy that year.[8] A gynecologist looks back at a random sample of medical records of patients who were prescribed the pill one year ago.

a. What are the mean and standard deviation of the distribution of sample proportions of women experiencing unplanned pregnancies?
b. The gynecologist takes an SRS of 200 records and finds that 14 women became pregnant within one year while taking the pill. How surprising

is this finding? Give the probability of finding 7% or more pregnant women in the sample.

c. How surprising would it be if the gynecologist found 16 pregnant women (8%)? 20 pregnant women (10%)?

CHAPTER 14 Introduction to Inference



BSIP/UIG/Getty Images

IN THIS CHAPTER WE COVER...

- Statistical estimation
- Margin of error and confidence level
- Confidence intervals for the mean μμ
- Hypothesis testing
- *P*-value and statistical significance
- Tests for a population mean
- Tests from confidence intervals

After we have selected a sample, we know the values recorded for the individuals in the sample. The usual reason for taking a sample is not to learn about the individuals in the sample, but rather to *infer* from the sample data some conclusion about the wider population that the sample represents.

STATISTICAL INFERENCE

Statistical inference provides methods for drawing conclusions about a population from sample data.

Because a different sample might lead to different conclusions, we can’t be certain that our conclusions are correct. Statistical inference uses the language of probability to say how trustworthy our conclusions are. This chapter introduces the two most common types of inference: *confidence intervals* for estimating the value of a population parameter and *hypothesis tests* to assess the evidence against a claim about a population parameter. Both types of inference are based on the sampling distributions of statistics. That is, both report probabilities that state what would happen if we used the inference method many times.

This chapter presents the main ideas of statistical inference. Chapter 15 will discuss these ideas in greater depth. The emphasis in both chapters is placed on basic reasoning, rather than on computations or applications. To make the reasoning as clear as possible, we consider the simplest of situations for statistical inference about a population mean. Here is the setting for our work in this chapter.

SIMPLE CONDITIONS FOR INFERENCE ABOUT A MEAN

1. We have an SRS from the population of interest. There is no nonresponse or other practical difficulty.
2. The quantitative variable we measure has a perfectly Normal distribution N(μ, σ)$N(\mu, \sigma)$ in the population.
3. We don’t know the population mean μ.μ. But we do know the population standard deviation σ.σ.

This setting is too simple and perfect to be of great practical use, *especially in the life sciences*. Nevertheless, the ideas we develop under these simple conditions can be adapted to more realistic situations, as we will see in later chapters.

STATISTICAL ESTIMATION

The National Health and Nutrition Examination Survey (NHANES) reports the weights and heights of Americans every few years. Each survey is based on a nationwide probability sample of the U.S. civilian noninstitutionalized population. What can we learn from the survey about the height of eight-year-old boys in America?

EXAMPLE 14.1 How tall are eight-year-old boys?

A recent NHANES reports that the mean height of a sample of 217 eight-year-old boys was $\bar{x} = 132.5$ centimeters (52.2 inches). On the basis of this sample, we want to estimate the mean height μ in the population of more than a million American eight-year-old boys.

Nick White/Getty Images

To match the "simple conditions," we will treat the NHANES sample as a perfect SRS of all American eight-year-old boys and the height in this population as having an exactly Normal distribution with standard deviation $\sigma = 10\text{ cm}$.[1]

We learned in Chapter 13 that the statistic $\bar{x}$ is an unbiased estimator of the parameter μ, but how far from μ could a given $\bar{x}$ be? Here is the reasoning of statistical estimation in a nutshell.

1. To estimate the unknown population mean μ, use the mean $\bar{x} = 132.5$ of the random sample. We don't expect $\bar{x}$ to be

exactly equal to μ,μ, so we want to say how accurate this estimate is.

2. We know the sampling distribution of x¯.$\bar{x}$. In repeated samples, x¯ $\bar{x}$ has the Normal distribution with mean μμ and standard deviation σ/n.$\sigma/\sqrt{n}$. So the sampling distribution of the average height x¯$\bar{x}$ of 217 eight-year-old American boys has standard deviation

   σn=10217=0.7 cm (rounded off)

   $$\frac{\sigma}{\sqrt{n}}=\frac{10}{\sqrt{217}}=0.7\,\text{cm (rounded off)}$$

3. The 95 part of the approximate 68–95–99.7 rule for Normal distributions says that x¯$\bar{x}$ is within 1.4 cm (that's two standard deviations) of the mean μμ in about 95% of all samples. That is, for 95% of all samples, 1.4 cm is the maximum distance separating x¯$\bar{x}$ and μ.μ. Therefore, if we estimate that the value of μμ is somewhere in the interval from x¯-1.4$\bar{x}-1.4$ to x¯+1.4,$\bar{x}+1.4$, we'll be right 95% of the times we take a sample. For this particular NHANES sample, this interval is

   x¯-1.4=132.5-1.4=131.1

   $$\bar{x}-1.4=132.5-1.4=131.1$$

   to

   x¯+1.4=132.5+1.4=133.9

   $$\bar{x}+1.4=132.5+1.4=133.9$$

   Because we got the interval 131.1 to 133.9 cm from a method that captures the population mean μμ 95% of the time, we say that we are *95% confident* that the mean height of all eight-year-old boys in the United States is some value in that interval, perhaps as low as 131.1 cm or as high as 133.9 cm.

RANGES ARE FOR STATISTICS?

Many people like to think that statistical estimates are exact. The Nobel prize–winning economist Daniel McFadden tells a story of his time on the Council of Economic Advisers. When presented with a range of forecasts for economic growth, President Lyndon Johnson replied: "Ranges are for cattle;

give me one number.”

The big idea is that the sampling distribution of $\bar{x}$ tells us how close to μ the sample mean $\bar{x}$ is likely to be. A confidence interval turns that information around to say how close to $\bar{x}$ the unknown population mean μ is likely to be.

EXAMPLE 14.2 Statistical estimation in pictures

Figure 14.1 summarizes what we learned in Chapter 13 about the sampling distribution of $\bar{x}$. Starting with the population, imagine taking many SRSs of 217 eight-year-old American boys. Perhaps one sample has mean height $\bar{x} = 132.5\ \text{cm}$, a second sample has mean $\bar{x} = 134.2\ \text{cm}$, a third has mean $\bar{x} = 131.8\ \text{cm}$, and so on. If we could collect all these sample means and display their distribution, we would get the Normal distribution with mean equal to the unknown value of the parameter μ and standard deviation 0.7 cm.

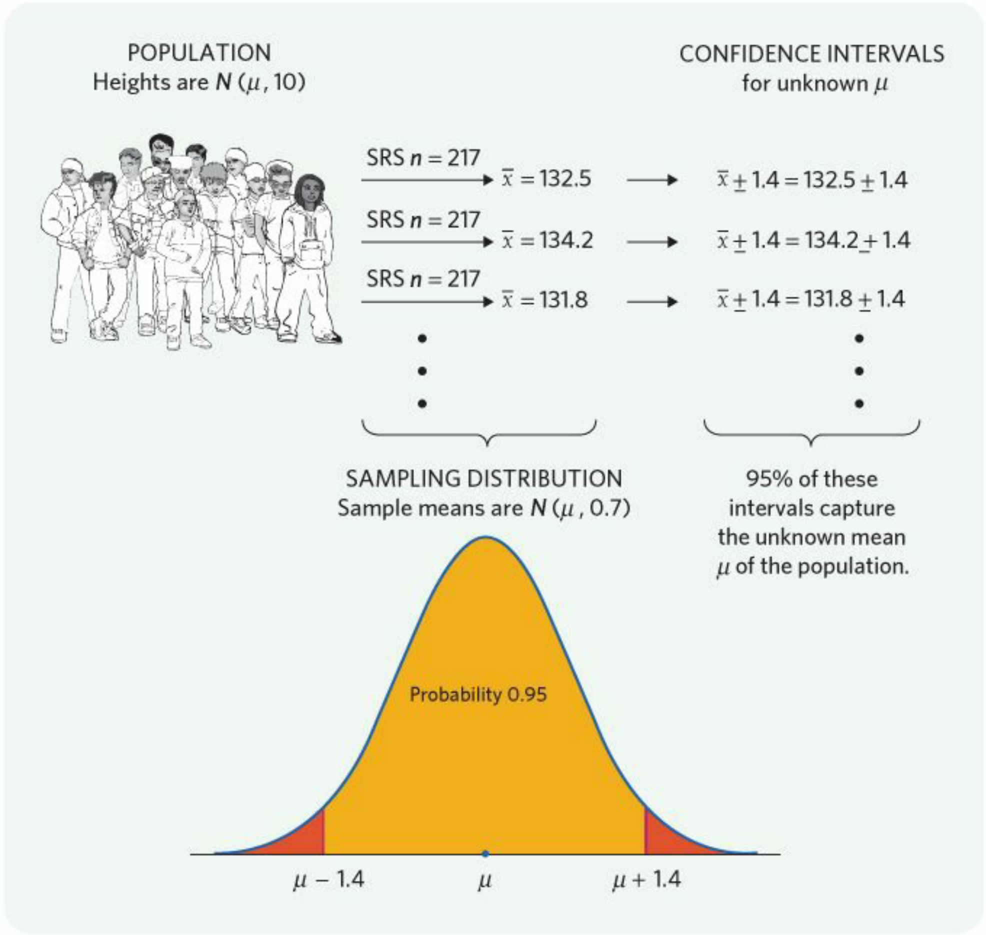
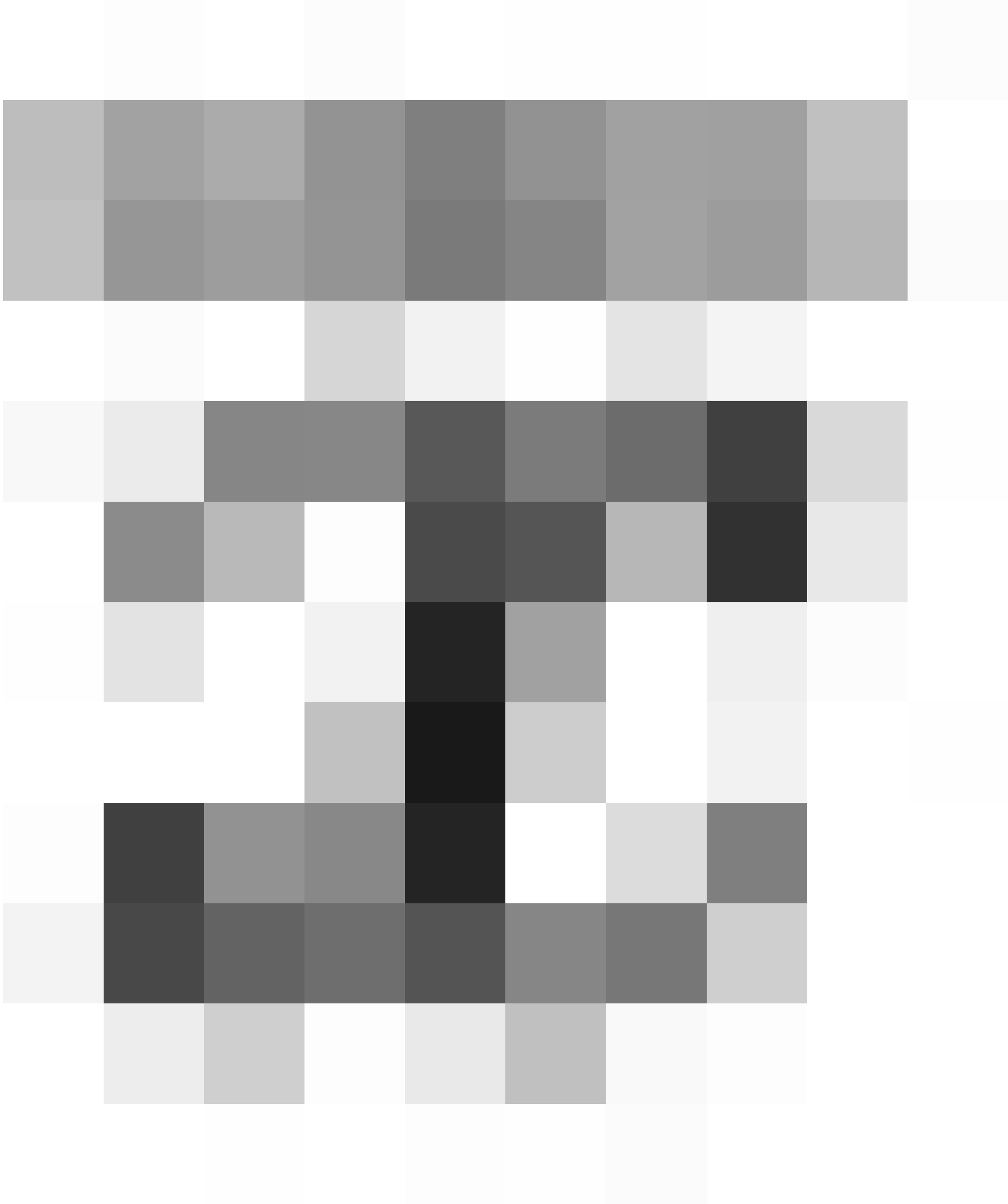

POPULATION
Heights are $N(\mu, 10)$
CONFIDENCE INTERVALS
for unknown μ
SRS $n = 217$
$\bar{x} = 132.5$
$\bar{x} \pm 1.4 = 132.5 \pm 1.4$
SRS $n = 217$
$\bar{x} = 134.2$
$\bar{x} \pm 1.4 = 134.2 \pm 1.4$
SRS $n = 217$
$\bar{x} = 131.8$
$\bar{x} \pm 1.4 = 131.8 \pm 1.4$
SAMPLING DISTRIBUTION
Sample means are $N(\mu, 0.7)$
95% of these intervals capture the unknown mean μ of the population.
Probability 0.95
$\mu - 1.4$
μ
$\mu + 1.4$

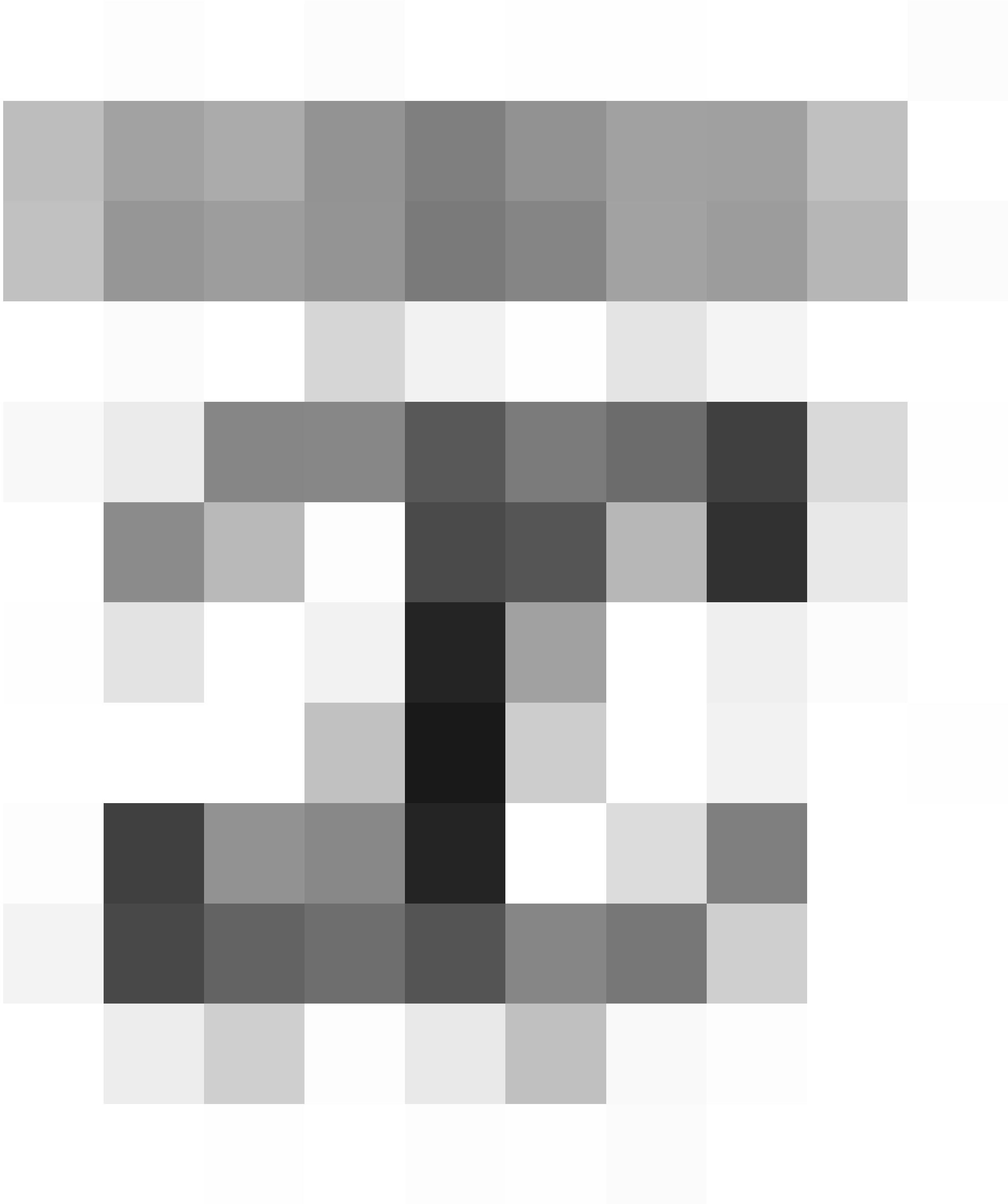
Figure 14.1

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 14.1 The sampling distribution of the mean height $\bar{x}$



of an SRS of 217 eight-year-old boys. In 95% of all samples, $\bar{x}$



lies within ± 1.4

± 1.4

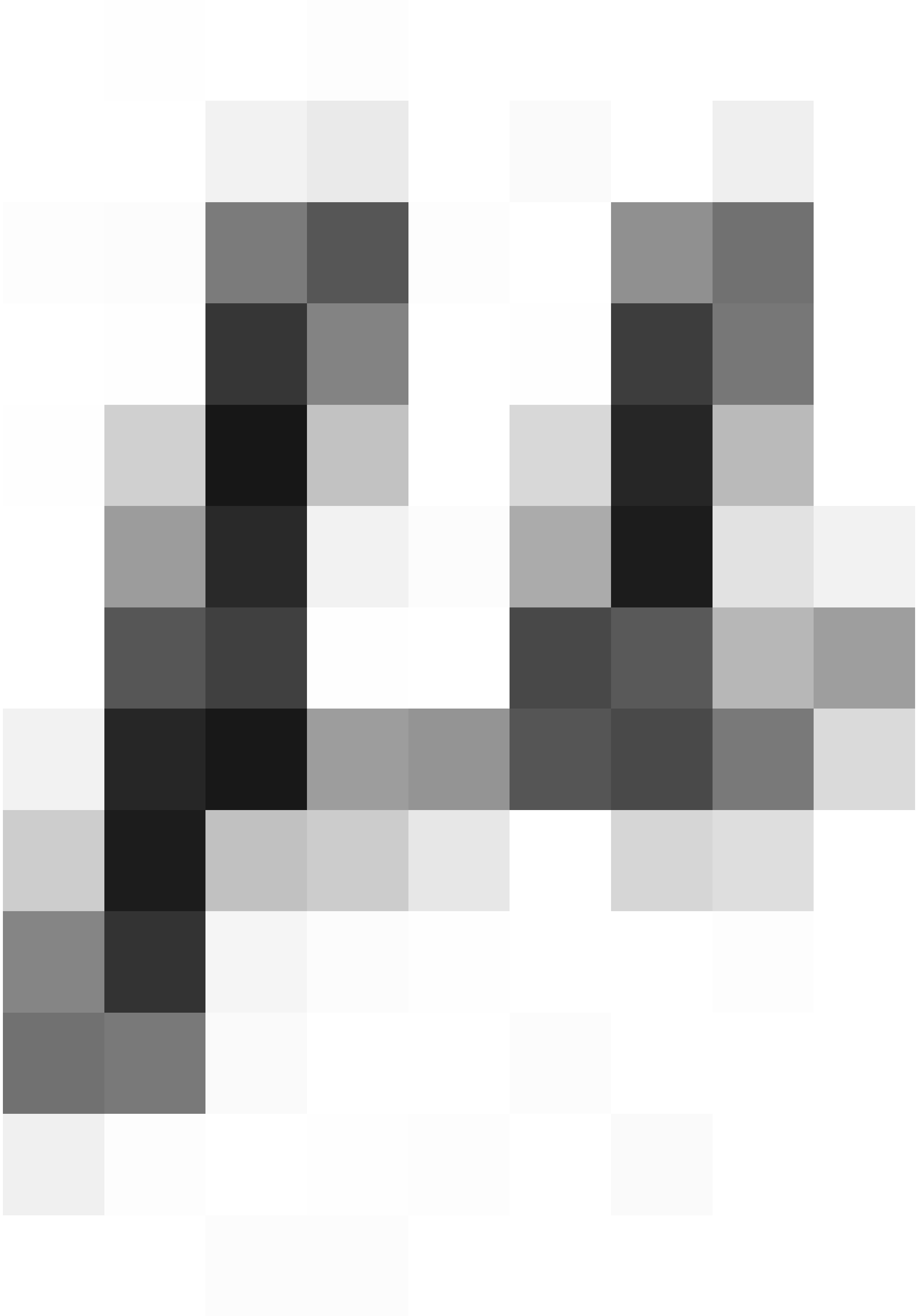
of the unknown population mean μ.

μ

To say that $\bar{x} \pm 1.4$

$$\bar{x} \pm 1.4$$

is a 95% confidence interval for the population mean μ



is to say that, in repeated samples, 95% of those intervals capture μ.

μ

The approximate 68–95–99.7 rule for Normal calculations says that in 95% of all samples, the sample mean x¯$\bar{x}$ is within 1.4 cm (two standard deviations) of the population mean μ.μ. Whenever this happens, the interval x¯±1.4$\bar{x} \pm 1.4$ captures the value of μ.μ. The formula x¯±1.4 $\bar{x} \pm 1.4$ gives an interval based on each sample; *95% of these intervals capture the unknown value of the population mean μ.*μ. That is, a statistical method using the interval x¯±1.4$\bar{x} \pm 1.4$ has a *95% success rate* in capturing within that interval the mean height μμ of all eight-year-old American boys.

It is important to understand that we don't need to take many SRSs to estimate μ.μ. Because the method has a 95% success rate, all we need is one SRS of 217 eight-year-old American boys to compute just one interval $\bar{x} \pm 1.4$. We are 95% *confident* that this particular interval contains the unknown value of the parameter μ. This interval of numbers between the values $\bar{x} - 1.4$ and $\bar{x} + 1.4$ is called a *95% confidence interval* for μ.

APPLY YOUR KNOWLEDGE

14.1 Girls' heights. Suppose that you measure the heights of an SRS of 400 eight-year-old American girls, a population with mean $\mu = 140$ cm and standard deviation $\sigma = 8$ cm. The mean $\bar{x}$ of the 400 heights will vary if you take repeated samples.

a. The sampling distribution of $\bar{x}$ is approximately Normal. It has mean $\mu = 140$ cm. What is its standard deviation?

b. Sketch the Normal curve that describes how $\bar{x}$ varies in many samples from this population. Mark the mean $\mu = 140$. According to the 68–95–99.7 rule, approximately 95% of all the values of $\bar{x}$ fall within ____________ cm of the mean. What is the missing number? Call it m for "margin of error." Shade the region from the mean minus m to the mean plus m on the axis of your sketch, as in Figure 14.1.

c. Whenever $\bar{x}$ falls in the region you shaded, the true value of the population mean, $\mu = 140$, lies in the interval between $\bar{x} - m$ and $\bar{x} + m$. Draw that interval below your sketch for one value of $\bar{x}$ inside the shaded region and one value of $\bar{x}$ outside the shaded region.

d. In what percent of all samples will the confidence interval $\bar{x} \pm m$ capture the true mean $\mu = 140$?

MARGIN OF ERROR AND CONFIDENCE LEVEL

Our confidence interval for the mean height of all eight-year-old boys, based on the NHANES sample, is 132.5±1.4.132.5 ± 1.4. Most confidence intervals take the form

$$\text{estimate} \pm \text{margin of error}$$

$\text{estimate} \pm \text{margin of error}$

The estimate (x¯=132.5 cm$\bar{x} = 132.5\text{ cm}$ in our example) is the center of the interval. It is our unbiased guess for the value of the unknown parameter, based on the sample data. The *margin of error* (in our example, 1.4 cm 1.4 cm) shows how accurate we believe our guess is, based on the variability of the estimate. We have a *95%* confidence interval because the interval catches the unknown value of the parameter (μμ in this case) with 95% probability based on the overall success rate of the method. That is, we are 95% confident that the unknown value of the parameter is a value within the confidence interval.

CONFIDENCE INTERVAL

A **level** *C* **confidence interval** for a parameter has two parts:

- An interval calculated from the data, usually of the form

$$\text{estimate} \pm \text{margin of error}$$

$\text{estimate} \pm \text{margin of error}$

  where the estimate is a sample statistic and the **margin of error** represents the accuracy of our guess for the parameter.

- A **confidence level** *C*, which gives the probability that the interval will capture the true parameter value in repeated samples. That is, the confidence level is the success rate for the method.

Users can choose the confidence level, which is usually set at 90% or higher because we want to be reasonably sure of our conclusions. It may be disconcerting at first to rely on a method that is not 100% reliable. However, we must keep in mind that the only way to be 100% confident about the unknown value of a parameter would be to record the entire population; this is rarely an option, especially in the life sciences.

INTERPRETING A CONFIDENCE INTERVAL

The confidence level is the success rate of the method that produces the interval. We don't know whether the 95% confidence interval from a particular sample is one of the 95% of intervals that capture μ or one of the unlucky 5% of intervals that miss.

To say that we are **95% confident** that the unknown value of μ is between 131.1 and 133.9 cm is shorthand for **"We got these numbers using a method that gives correct results 95% of the time."**

Figure 14.2 uses computer simulations (as we did in Chapter 13) to illustrate the ideas of confidence interval and confidence level. In this simulation, each SRS consists of $n=15$ observations randomly selected from a given population. Both the population distribution and the sampling distribution of $\bar{x}$ appear at the top of the figure (in solid and dotted lines, respectively). The population mean μ is at the center of both distributions. The 95% confidence intervals derived from 25 different SRSs appear underneath. The center of each interval is at $\bar{x}$ (marked by a dot) and, therefore, varies from sample to sample. The arrows on either side of the dot span the confidence interval. All except 1 of these 25 intervals contain the true value of μ. In a very large number of samples, 95% of the confidence intervals would contain μ.

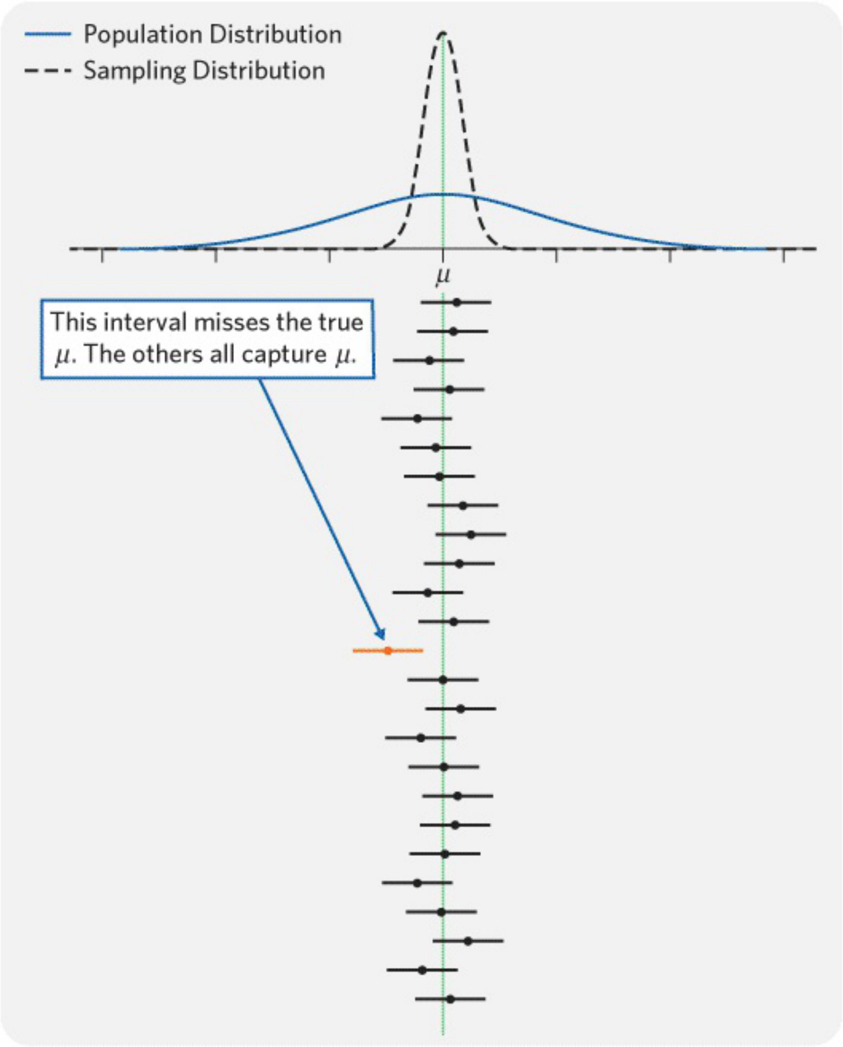


Figure 14.2

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 14.2 Twenty-five SRSs from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that contains the population

mean μ.μ

The *Confidence Intervals* applet lets you perform such simulations and watch confidence intervals from one sample after another capture or fail to capture the true parameter. Notice that two things influence the size of the simulated confidence intervals: the confidence level *C* and the sample size *n*. The applet also allows you to display the *n* individual observations making up an SRS by clicking on the interval.

APPLY YOUR KNOWLEDGE

14.2 Confidence intervals in action. The idea of an 80% confidence interval is that the interval captures the true parameter value in 80% of all samples. That's not a high enough confidence level for practical use, but 80% hits and 20% misses make it easy to see how a confidence interval behaves in repeated samples from the same population. Go to the *Confidence Intervals* applet.

APPLET

a. Pick a sample size—say, n=15.$n = 15$. Set the confidence level to

80%. Click “Sample” to choose an SRS and calculate the confidence interval. Do this 10 times to simulate 10 SRSs with their 10 confidence intervals. How many of the 10 intervals captured the true mean μ?μ? How many missed?

b. You see that we can’t predict whether the next sample will hit or miss. The confidence level, however, tells us what percent will hit in the long run. Reset the applet and click “Sample 25” to get the confidence intervals from 25 SRSs. How many hit? Keep clicking “Sample 25” and record the percent of hits among 100, 500, and 1000 SRSs. Even 1000 samples is not truly “the long run,” but we expect the percent of hits in 1000 samples to be fairly close to the confidence level, 80%.

14.3 Spending on food. A 2012 Gallup survey of a random sample of 1014 American adults indicates that American families spend, on average, \$151 per week on food. The report further states that, with 95% confidence, this estimate has a margin of error of ±\$7.±\$7.

a. This confidence interval is expressed in the following form: “estimate±margin of error.estimate ± margin of error.” What is the range of values (lower bound, upper bound) that corresponds to this confidence interval?

b. What is the parameter captured by this confidence interval? What does it mean to say that we have “95% confidence” in this interval?

CONFIDENCE INTERVALS FOR THE MEAN μμ

We computed a 95% confidence interval for the mean height μμ of eight-year-old American boys in the form of x¯±2σ/n.$\bar{x} \pm 2\sigma/\sqrt{n}.$ This approach relies on the approximate 68–95–99.7 rule for Normal distributions, which tells us that the central 95% of the Normal sampling distribution of x¯$\bar{x}$ spreads out 2 standard deviations in both directions from the mean μ.$\mu.$ We can use this reasoning to compute a 68% confidence interval for μμ in the form of x¯±σ/n,$\bar{x} \pm \sigma/\sqrt{n},$ or a 99.7% confidence interval for μμ in the form of x¯±3σ/n.$\bar{x} \pm 3\sigma/\sqrt{n}.$

In fact, we can choose any confidence level *C* by finding the proper value to catch the central area *C* under the Normal sampling distribution. Figure 14.3 shows how the central area *C* under a standard Normal curve is marked off by two points z*z^* and -z*.$-z^*.$ Numbers like z*z^* that mark off specific areas of a distribution are called **critical values** of that distribution. Here are the Normal critical values for the most common confidence levels:

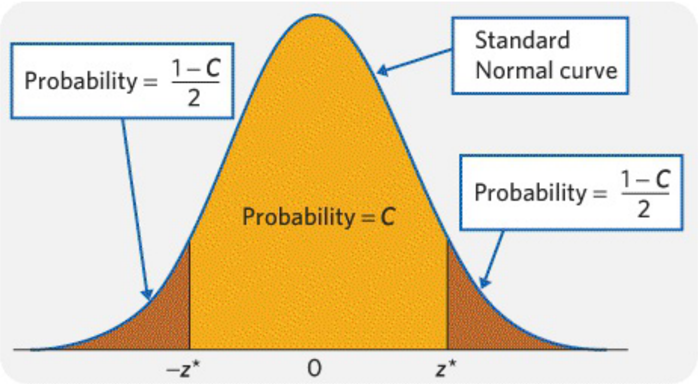


Figure 14.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 14.3 The critical value z* is the number that catches central probability C under a standard Normal curve between −z* and z*.

Confidence level *C*	90%	95%	99%
Critical value z*z^*	1.645	1.960 (≈2)(≈ 2)	2.576

These values are easily obtained with software functions for inverse Normal calculations, as seen in Chapter 11. (You can also find these and a few more critical values in the bottom row labeled z^* of Table C in the back of the book.) Notice that for $C = 95\%$, the critical value is actually $z^* = 1.960$. This is a bit more precise than the approximate value $z^* = 2$ based on the 68–95–99.7 rule.

If we start at the sample mean $\bar{x}$ and move outward by z^* standard deviations on either side, we get an interval that contains the population mean μ in a proportion *C* of all samples. This interval is

from x‾-z*σn to x‾+z*σn

$$\text{from } \bar{x} - z^* \frac{\sigma}{\sqrt{n}} \text{ to } \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

or

x‾±z*σn

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

It is a level *C* confidence interval for μ.

CONFIDENCE INTERVAL FOR THE MEAN OF A NORMAL POPULATION

Draw an SRS of size n from a Normal population having unknown mean μ and known standard deviation σ. A level *C* **confidence interval for** μ is

x‾±z*σn

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Any technology with basic statistical inference capabilities will perform all necessary computations and produce the lower and upper boundaries of the confidence interval.

This and other types of confidence intervals for specific parameters that we will describe in later chapters take the following form:

unbiased±(critical×(standard deviation of theestimatevalue)distribution of the estimate)

$$\begin{array}{ccccc}\text{unbiased} & \pm & (\text{critical} & \times & (\text{standard deviation of the} \\ \text{estimate} & & \text{value}) & & \text{distribution of the estimate})\end{array}$$

The steps in finding confidence intervals mirror the overall four-step process for organizing statistical problems.

CONFIDENCE INTERVALS: THE FOUR-STEP PROCESS

STATE: What is the practical question that requires estimating a parameter?

PLAN: Identify the parameter and choose a level of confidence.

SOLVE: Carry out the work in two phases:

1. **Check the conditions** for the interval you plan to use.
2. Obtain the **confidence interval**.

CONCLUDE: Return to the practical question to describe your results in this setting.

EXAMPLE 14.3 IQ scores

STATE: We are interested in the mean IQ score of seventh-grade girls in a Midwest school district. Here are the scores for 31 randomly selected seventh-grade girls in the district:[2]

114	100	104	89	102	91	114	114	103	105	
108	130	120	132	111	128	118	119	86	72	
111	103	74	112	107	103	98	96	112	112	93

PLAN: We will estimate the mean IQ score μμ for all seventh-grade girls in this Midwest school district by giving a 95% confidence interval.

SOLVE: **Check the conditions for inference.** Here are the “simple conditions”:

1. **SRS:** The girls are an SRS from the population of all seventh-grade girls in the school district.
2. **Normal distribution:** IQ score is a quantitative variable taking

numerical values that are typically rounded to the nearest integer when provided. We expect from experience that IQ scores in a homogeneous population with similar developmental age will follow approximately a Normal distribution. We can’t look at the population, but we can examine the sample. Figure 14.4 shows a dotplot of the data. Its shape is slightly left-skewed, but there are no extreme outliers. Thus, we have no reason to doubt that the population distribution is close to Normal. Additionally, the central limit theorem tells us that a sample size of 31 is large enough to overcome moderate departures from Normality for performing inference.

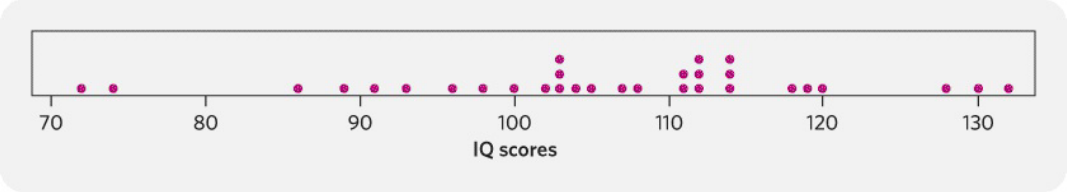


Figure 14.4

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 14.4 **Dotplot of the IQ scores of 31 seventh-grade girls.**

3. **Known** σ:σ **:** Most IQ tests are designed so that $\sigma=15$,$\sigma = 15$, so it is not unrealistic to assume that we know that $\sigma=15$$\sigma = 15$ for our calculations.

SOLVE: Obtain the confidence interval. Using statistical software and inputting the raw data, we obtain a 95% confidence interval of (100.56, 111.12) for the mean IQ score $\mu$$\mu$ in this population.

CONCLUDE: We are 95% confident that the mean IQ score for all seventh-grade girls in this Midwest school district is between 100.6 and 111.1.

The computations performed by any statistical software in this example involve the following steps:

- Sample mean: $\overline{x}=105.84$$\bar{x} = 105.84$
- Critical value for a 95% confidence level: $z^*=1.960$$z^* = 1.960$
- Margin of error: $m=z^*\sigma/n=(1.96)(15)/31=5.28$

  $m = z^*\sigma/\sqrt{n} = (1.96)\,(15)\,/\sqrt{31} = 5.28$

- 95% confidence interval for $\mu : \bar{x} \pm m = 105.84 \pm 5.28$, or $(100.56, 111.12)$

In this introductory chapter, we work with examples for which the “simple conditions” are likely to be satisfied. More broadly, inference methods are often used when conditions such as SRS and Normal population are not exactly met. In addition, knowing σ is the exception rather than the rule, so it is not generally realistic to suppose that we know σ. In Chapter 17, you will learn about an approach to do away with the need to know σ. Chapter 15 and the later chapters on each inference topic will give you a basis to judge whether it is safe to do inference in specific instances when the simple conditions are not met.

APPLY YOUR KNOWLEDGE

14.4 Critical value and confidence level. This section provided the Normal critical values z^* for three commonly used confidence levels (90%, 95%, and 99%). Explain briefly why the critical value must be larger for a larger confidence level.

14.5 Pharmaceutical production. A manufacturer of pharmaceutical products analyzes each batch of a product to verify the concentration of the active ingredient. The chemical analysis is not perfectly precise. In fact, repeated measurements follow a Normal distribution with mean μ equal to the true concentration and with standard deviation $\sigma = 0.0068$ grams per liter (g/l). Three analyses of one batch give concentrations of 0.8403, 0.8363, and 0.8447 g/l. To estimate the true concentration, give a 95% confidence interval for μ. Follow the four-step process as illustrated in Example 14.3.

HYPOTHESIS TESTING

Confidence intervals are one of the two most common types of statistical inference. Use a confidence interval when your goal is to estimate a population parameter. Another common type of inference, called *hypothesis tests* (or sometimes significance tests), has a different goal: to assess the strength of the evidence provided by data *against* some claim concerning a population. Here is an example.

EXAMPLE 14.4 Inorganic phosphorus levels in the elderly

Phosphorus is a mineral that is essential to a whole range of metabolic processes as well as to bone stiffening. Levels of inorganic phosphorus in the blood are known to vary among adults Normally with mean 1.2 millimoles per liter (mmol/l) and standard deviation 0.1 mmol/l.

adriaticfoto/Shutterstock.com

A study was conceived to examine inorganic phosphorus blood levels in older individuals to see if it decreases with age, since we know that older people are more prone to bone fractures. Here are data from a retrospective chart review of 12 men and women between the ages of 75 and 79 years:[3]

1.26	1.39	1.00	1.00	1.00	1.10
0.87	1.23	1.19	1.29	1.03	1.18

Some of these values are above the adult population mean and some are below it. The average inorganic phosphorus level in the study is given by

> the sample mean $\bar{x} = 1.128\ \text{mmol/l}$. Are these data good evidence that, on average, inorganic phosphorus levels among people age 75 to 79 are lower than in the whole adult population?

A statistical hypothesis test starts with a careful statement of the claims we want to examine regarding some property of one or more populations. In Example 14.4 we considered the competing claims that the mean inorganic phosphorus blood level μ in the elderly population is either 1.2 mmol/l (no difference from the adult mean of 1.2 mmol/l) or less than 1.2 mmol/l (lower than the mean of the adult population, a consequence of aging). These competing claims are formally expressed as a null hypothesis and an alternative hypothesis.

NULL AND ALTERNATIVE HYPOTHESES

The specific claim about a population of interest tested by a hypothesis test is called the **null hypothesis.** A hypothesis test is designed to assess the strength of the evidence *against* the null hypothesis. Usually the null hypothesis is a statement of “no effect” or “no difference.”

The more general claim about the population that we are trying to find evidence *for* is the **alternative hypothesis.**

- The alternative hypothesis is **one-sided** if it states that a parameter is *larger than* or that it is *smaller than* the null hypothesis value.
- It is **two-sided** if it states that the parameter is *different from* the null value (it could be either smaller or larger).

We denote the null hypothesis as H_0 and the alternative hypothesis as H_a (the notation H_1 is often used as well). *Hypotheses always refer to a population, not to a particular outcome. Be sure to state* H_0 *and* H_a *in terms of population parameters.*

EXAMPLE 14.5 Inorganic phosphorus: hypotheses

The study in Example 14.4 is seeking evidence *for* a lower inorganic phosphorus level in the elderly, on average. The null hypothesis says that “there is no difference from the adult mean of 1.2 mmol/l,” on average, in a large population of elderly individuals between the ages of 75 and 79 years. The alternative hypothesis says that “their mean is lower than 1.2 mmol/l.” So the hypotheses are

H0:μ=1.2Ha:μ<1.2

$$\begin{aligned} H_0 &: \mu = 1.2 \\ H_a &: \mu < 1.2 \end{aligned}$$

The alternative hypothesis is *one-sided* because we are interested only in whether the elderly have a *lower* inorganic phosphorus level, on average. This choice of alternative hypothesis is based on the biological hypothesis that some physiological processes such as mineral absorption may become less efficient with age, leading to lower blood levels.

EXAMPLE 14.6 Manufacturing aspirin tablets

A pharmaceutical company manufactures aspirin tablets that are sold with the label “active ingredient: aspirin 325 mg.” No production, even machine-made, is ever perfect, and individual tablets do vary a little bit in their actual aspirin content. A small amount of variation is acceptable provided that, on average, the whole production has mean μ=325 mg.

$\mu = 325\ \mathrm{mg}.$

Glowimages/Getty Images

The parameter of interest is the mean aspirin content per tablet, μ. The null hypothesis says that the population of all aspirin tablets produced by this manufacturer has mean $\mu = 325$ mg:

$$H_0: \mu = 325 \text{ mg}$$

$H_0 : \mu = 325 \text{ mg}$

Proper dosage of prescription medicine is important, and deviations from the expected dose distribution in either direction (too much or too little aspirin per tablet) would need to be identified and then, of course, remedied. The alternative hypothesis is therefore *two-sided:*

$$H_a: \mu \neq 325 \text{ mg}$$

$H_a : \mu \neq 325 \text{ mg}$

Statistical hypotheses should reflect a legitimate inquiry grounded in existing knowledge on the topic, not some arbitrary choice. In particular, *the hypotheses should express the hopes or suspicions we have* ***before*** *we see the data. It is cheating to first look at the data and then frame hypotheses to fit what the data show.* Thus, the fact that a random sample of aspirin tablets in Example 14.6 gave a sample average $\bar{x}$ larger than the population mean μ should not influence our choice of H_a. When you do not have a specific direction firmly in mind in advance, you must use a two-sided alternative.

APPLY YOUR KNOWLEDGE

14.6 Stating hypotheses. In planning a study of the birth weights of babies whose mothers did not see a doctor before delivery, a researcher states the hypotheses as follows:

$$H0:\bar{x}=1000\text{ grams}Ha:\bar{x}<1000\text{ grams}$$

$$\begin{aligned} H_0 : \bar{x} &= 1000\text{ grams} \\ H_a : \bar{x} &< 1000\text{ grams} \end{aligned}$$

What’s wrong with this formulation?

14.7 Anemia. Hemoglobin is a protein in red blood cells that carries oxygen from the lungs to body tissues. People with less than 12 grams of hemoglobin per deciliter of blood (g/dl) are anemic. A public health official in Jordan suspects that the mean μμ for all children in Jordan is less than 12 g/dl because of inadequate health conditions. The official measures hemoglobin levels in a sample of 50 children. Define the parameter examined and state the appropriate null and alternative hypotheses.

Laguna Design/Getty Images

14.8 Arsenic contamination. Arsenic is a compound that occurs naturally in very low concentrations. Arsenic blood concentrations in healthy individuals are Normally distributed, with mean μ=3.2 micrograms per deciliter(μg/dl)$\mu = 3.2\text{ micrograms per deciliter}(\mu\text{g/dl})$ and standard deviataion σ=1.5 μg/dl.$\sigma = 1.5\,\mu\text{g/dl}$.[4] Some areas are known to have naturally elevated concentrations of arsenic in the ground and water supplies. We take a random sample of adults residing in a high-arsenic area and ask whether the data provide good evidence that the mean blood arsenic concentration in this population is elevated compared with the 3.2 mg/dl mean of the population of healthy individuals. Define the parameter examined and state the appropriate null and alternative hypotheses.

P-VALUE AND STATISTICAL SIGNIFICANCE

The idea of making a claim (H0)(H_0) so as to evaluate evidence *against* it might seem odd at first. The approach is not unique to hypothesis tests, though. Think of a criminal trial: The defendant is “innocent until proven guilty.” That is, the null hypothesis is innocence, and the prosecution must try to provide convincing evidence against this hypothesis. That’s exactly how statistical tests work, although in statistics we deal with evidence provided by data and use a probability to say how strong the evidence is.

In Example 14.4 we sought evidence that the mean blood level of inorganic phosphorus in people age 75 to 79 *is less than* 1.2 mmol/l (Ha), $1.2\,\text{mmol/l}\,(H_a)$, so the specific claim we test is that the mean for people age 75 to 79 *is* 1.2 mmol/l (H0).$1.2\,\text{mmol/l}\,(H_0)$. If the claim that μ=1.2$\mu = 1.2$ is true, the sampling distribution of x¯ $\bar{x}$ from 12 individuals age 75 to 79 is Normal with mean μ=1.2$\mu = 1.2$ and standard deviation

σn=0.112=0.0289

$$\frac{\sigma}{\sqrt{n}} = \frac{0.1}{\sqrt{12}} = 0.0289$$

Figure 14.5 shows this sampling distribution. We can judge whether any observed x¯ $\bar{x}$ is surprising by locating it on this distribution.

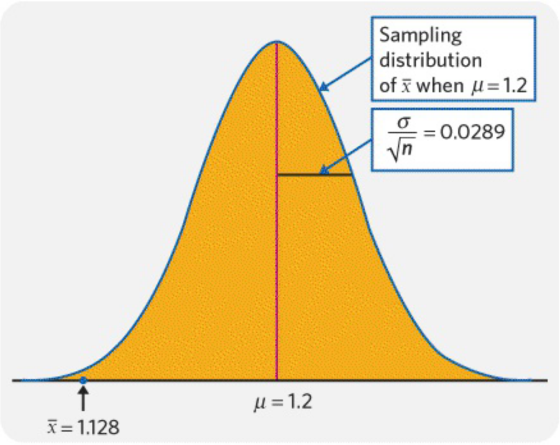


Figure 14.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 14.5 If people age 75 to 79 do not differ from the whole adult population in their blood levels of inorganic phosphorus, the mean inorganic phosphorus level x¯

$\bar{x}$ for a random sample of 12 individuals age 75 to 79 will have this sampling distribution. The actual result for this group is $\bar{x}=$ $\bar{x} =$ 1.128. That point is so far out on the Normal curve that it is good evidence that people age 75 to 79 do have inorganic phosphorus levels lower than 1.2 mmol/l.

- A sample mean $\bar{x}$ $\bar{x}$ close to the hypothesized population mean, within the central bulk of the hypothesized sampling distribution, would indicate that such a sample mean could easily occur just by chance when the population mean is $\mu=1.2$ $\mu = 1.2$
- In contrast, a sample mean $\bar{x}$ $\bar{x}$ located on the edges of the hypothesized sampling distribution would indicate that this outcome is somewhat unlikely to occur just by chance when the population mean is $\mu=1.2$. $\mu = 1.2$.

The idea here is that *observing an outcome that would rarely happen if a hypothetical claim was true is good evidence that this claim is not true.* In Example 14.4, the sample of 12 individuals age 75 to 79 produced

$\bar{x} = 1.128$ mmol/l. That's far out on the Normal curve in Figure 14.5—so far out that *an observed value this small or even smaller would rarely occur just by chance if the true population mean μ was 1.2 mmol/l.* The observed sample mean is good evidence that the true value of μ is, in fact, less than 1.2 mmol/l—that is, that the mean inorganic phosphorus blood level of elderly individuals between the ages of 75 and 79 years is lower than the known mean of 1.2 mmol/l in the general population.

The probability that measures the strength of the evidence against a null hypothesis is called a *P-value.* Statistical tests generally work as follows.

TEST STATISTIC AND *P*-VALUE

A **test statistic** calculated from the sample data measures how far the data diverge from the null hypothesis H_0. Large values of the statistic show that the data are far from what we would expect if H_0 was true.

The probability, computed assuming that H_0 is true, that the test statistic would take a value at least as extreme as (in the direction of H_a) as that actually observed is called the ***P*-value** of the test. The smaller the *P*-value, the stronger the evidence against H_0 provided by the data.

Small *P*-values are evidence against H_0, because they say that the observed result would be unlikely to occur if H_0 was true. That is, small *P*-values indicate that the data are inconsistent with H_0, so we should reject H_0. We then say that the results are **statistically significant.**

P-values that are not small, in contrast, are inconclusive: They fail to provide evidence against H_0, so H_0 cannot be rejected. *Failing to find evidence against* H_0 *means only that the data are not inconsistent with* H_0, *which does not at all imply that* H_0 *is true.*

Statistical software will give you the *P*-value of a test when you enter your null and alternative hypotheses and your data. Your most important task, then, is to understand what a *P*-value says.

EXAMPLE 14.7 Inorganic phosphorus: one-sided *P*-value

The study of inorganic phosphorus levels in the elderly in Example 14.5 tests the following hypotheses:

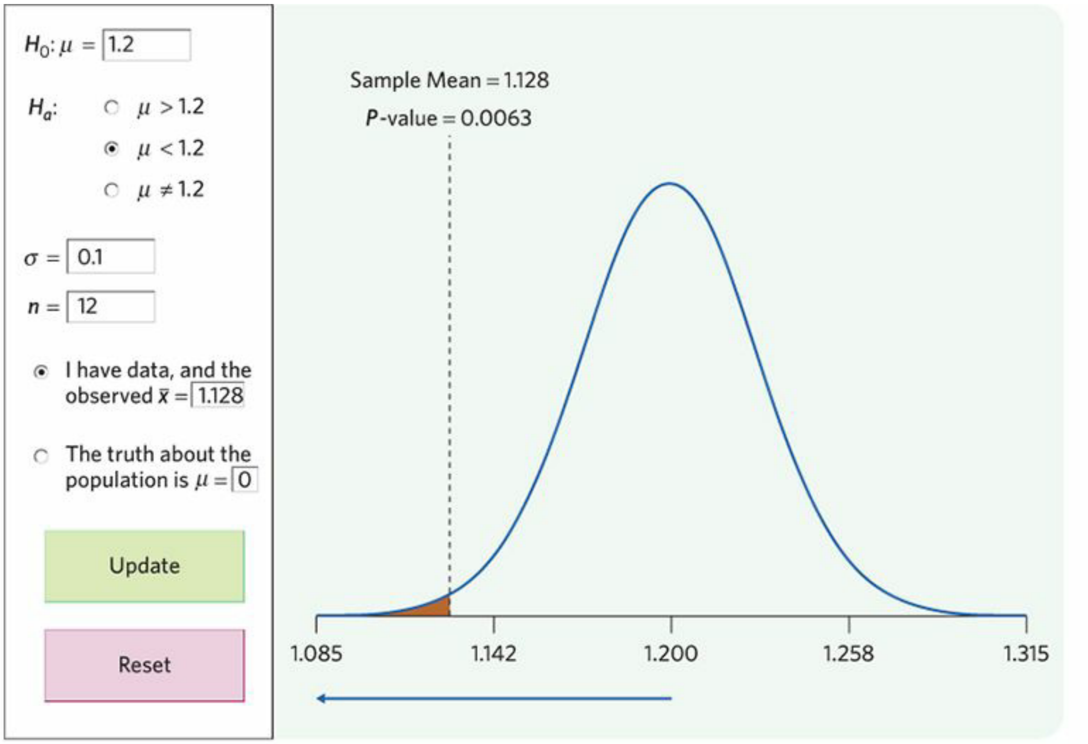
H0:μ=1.2Ha:μ<1.2

$$H_0 : \mu = 1.2$$
$$H_a : \mu < 1.2$$

Because the alternative hypothesis says that μ<1.2,$\mu < 1.2,$ values of x¯$\bar{x}$ less than 1.2 favor HaH_a over H0.$H_0.$ The 12 individuals age 75 to 79 had a mean inorganic phosphorus level of x¯=1.128.$\bar{x} = 1.128.$ *The P-value is the probability of getting an* x¯$\bar{x}$ *as small as 1.128 or smaller when the null hypothesis is really true.*

We act for now as if the "simple conditions" listed on page 348 are true: We have a perfect SRS from an exactly Normal population with standard deviation σσ known to us. The *P-Value of a Test of Significance* applet automates the work of finding *P*-values under the simple inference conditions. We enter the information for Example 14.5 into the applet—hypotheses, σ,$\sigma,$ n,$n,$ and x¯$\bar{x}$—then we click "Update." Figure 14.6 shows the applet output. It displays the *P*-value as an area under a Normal curve and its corresponding value P=0.0063.$P = 0.0063.$ (Note that this value is a simple Normal computation like the ones we saw in Chapter 11.)

APPLET


$H_0: \mu = 1.2$

H_a: ○ $\mu > 1.2$ ● $\mu < 1.2$ ○ $\mu \neq 1.2$

$\sigma = 0.1$

$n = 12$

● I have data, and the observed $\bar{x} = 1.128$

○ The truth about the population is $\mu = 0$

Sample Mean = 1.128

P-value = 0.0063

Figure 14.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 14.6 Output from the *P-Value of a Test of Significance* applet for Example 14.8. The *P*-value for the one-sided test, 0.0063, is the area under the Normal curve to the left of x¯=1.128.

$$\bar{x} = 1.128.$$

The small *P*-value says that we would rarely observe a sample mean inorganic phosphorus level of 1.128 or lower if H0 H_0 was true. This is strong evidence against H0 H_0 and in favor of the alternative Ha:μ<1.2. $H_a : \mu < 1.2.$ We conclude that the data provide strong, statistically significant evidence (P=0.0063 $P = 0.0063$) that the mean inorganic phosphorus level in the population of adults age 75 to 79 is less than 1.2 mmol/l.

The alternative hypothesis sets the direction that counts as evidence against H0. H_0. In Example 14.8, only smaller values count because the alternative is one-sided on the low side. If the alternative is two-sided, both directions count.

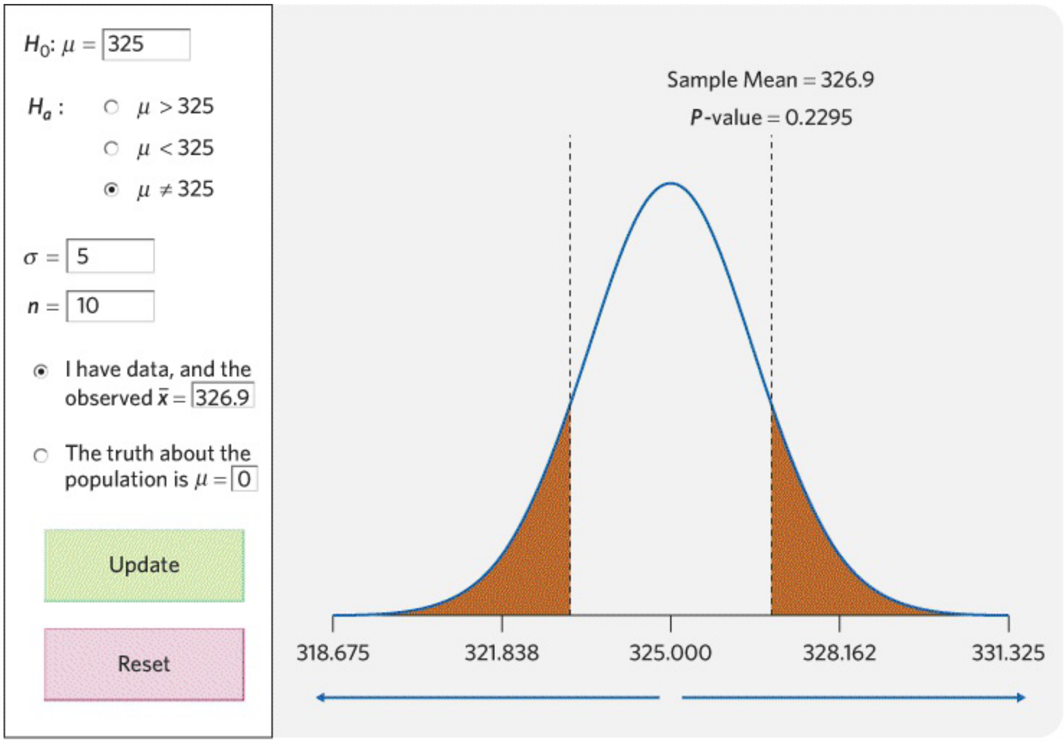
EXAMPLE 14.8 Aspirin tablets: two-sided *P*-value

Suppose that we know that the aspirin content of the tablets described in Example 14.6 follows a Normal distribution with standard deviation $\sigma = 5$ mg. If the manufacturing process is well calibrated, the mean aspirin content of all tablets is $\mu = 325$ mg. This is our null hypothesis. The alternative hypothesis says simply that “the mean of all tablets is not 325 mg.”

$$
\begin{aligned}
H_0 &: \mu = 325\\
H_a &: \mu \neq 325
\end{aligned}
$$

Data from a random sample of 10 aspirin tablets give $\bar{x} = 326.9$ mg. *Because the alternative is two-sided, the P-value is the probability of getting an* $\bar{x}$ *at least as far from the hypothesized* $\mu = 325$ *in either direction as the observed* $\bar{x} = 326.9$.

Enter the information for this example into the *P-Value of a Test of Significance* applet and click “Update.” Figure 14.7 shows the applet output as well as the information we entered. The *P*-value is the sum of the two shaded areas under the Normal curve: $P = 0.2295$. Values as far from 325 as $\bar{x} = 326.9$ (in either direction) would happen 23% of the time if the true population mean was $\mu = 325$. An outcome that would occur so often, if H_0 was true, is not good evidence against H_0.


H_0: μ = 325

H_a:
○ $\mu > 325$
○ $\mu < 325$
● $\mu \neq 325$

σ = 5

n = 10

● I have data, and the observed $\bar{x}$ = 326.9

○ The truth about the population is μ = 0

Figure 14.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 14.7 Output from the *P-Value of a Test of Significance* applet for Example 14.9. The *P*-value for the two-sided test, 0.2295, is twice the value of the area under the Normal curve to the right of x¯=326.9.

$$\bar{x} = 326.9.$$

We conclude that the data fail to provide statistically significant evidence (P=0.2295)$(P = 0.2295)$ that the mean aspirin content of all manufactured tablets is not 325 mg.

The conclusion of Example 14.8 is *not* that H0H_0 is true. The study looked for evidence against H0:μ=325$H_0 : \mu = 325$ and failed to find strong enough evidence against this hypothesis; *that is all we can say*. The mean μμ for the population of all aspirin tablets could, for instance, be a value close but not exactly equal to 325 mg.

In Examples 14.7 and 14.8, respectively, we decided that a *P*-value of

0.0063$0.0063$ was strong evidence against the null hypothesis and that a *P*-value of 0.2295$0.2295$ did not give convincing evidence against H0.H_0. There is no rule for how small a *P*-value we should require to reject H0H_0—it's a matter of judgment and depends to some extent on the specific circumstances.

Nonetheless, when reporting the *P*-value of a test, we can compare it with some fixed values that are in common use as standards for evidence againstH0.H_0. The most common fixed values are 0.05 and 0.01. If P≤0.05,$P \leq 0.05$, there is no more than 1 chance in 20 that a sample would give evidence this strong just by chance when H0H_0 is actually true. If P≤0.01,$P \leq 0.01$, we have a result that in the long run would happen no more than once per 100 samples when H0H_0 is true. These fixed standards for *P*-values are called *significance levels*. We use the Greek letter alpha, α,α, to stand for a significance level.

SIGNIFICANCE LEVEL

The **significance level,** α,α, is an arbitrary threshold that can be used when reporting whether a *P*-value is statistically significant.

If the *P*-value is as small as or smaller than α,α, we say that the data are **statistically significant at level** α.α.

The significance level αα can be chosen in advance to avoid human bias when deciding whether a *P*-value is statistically significant. It can also be used to describe statistical results. For example, statistical significance at level 0.01 can be expressed by the statement "The results were statistically significant (P<0.01$P < 0.01$)." Here *P* stands for the *P*-value. *Because the P-value assesses the strength of the evidence against* H0,H_0, *its order of magnitude matters more than its exact value*. For example, the *P*-values P=0.03$P = 0.03$ and P=0.0004$P = 0.0004$ are both significant at the α=0.05$\alpha = 0.05$ level, but P=0.0004$P = 0.0004$ offers stronger evidence against H0.H_0. It is also important to remember that *"significant" in the statistical sense does not mean "important." It means simply "not likely to happen just by chance because of random variations from sample to* *sample, if the null hypothesis was true.”*

APPLY YOUR KNOWLEDGE

14.9 Anemia, continued. Go back to the anemia study of Exercise 14.7. Suppose that the “simple conditions” hold: The 50 children are an SRS from all Jordanian children, and the hemoglobin level in this population follows a Normal distribution with standard deviation $\sigma=1.6$ g/dl.$\sigma = 1.6\text{ g/dl}.$

a. We seek evidence *against* the claim that $\mu=12$.$\mu = 12.$ What is the sampling distribution of $\bar{x}$ $\bar{x}$ in many samples of size 50 if, in fact, $\mu=12$?$\mu = 12?$ Make a sketch of the Normal curve for this distribution. (Sketch a Normal curve, then mark the axis using what you know about locating the mean and standard deviation on a Normal curve.)

b. The sample mean was $\bar{x}=11.3$ g/dl.$\bar{x} = 11.3\text{ g/dl}.$ Mark this outcome on the sampling distribution. Enter the hypotheses, n,$n,$ σ,$\sigma,$ and $\bar{x}$ $\bar{x}$ for the Jordan study into the *P-Value of a Test of Significance* applet. What is the *P*-value for this study? What do you conclude?

APPLET

c. A different study of 50 children from another country finds a sample mean $\bar{x}=11.8$ g/dl.$\bar{x} = 11.8\text{ g/dl}.$ Use the applet to obtain the *P*-value for this country. What do you conclude?

d. Explain briefly why these *P*-values tell us that one outcome is strong evidence against the null hypothesis and that the other outcome is not.

e. Is the outcome for either study statistically significant at the $\alpha=0.05$$\alpha = 0.05$ level? Explain briefly why reporting the actual *P*-value is more relevant to your conclusion than simply reporting whether the findings are statistically significant at the $\alpha=0.05$ $\alpha = 0.05$ level.

14.10 Arsenic contamination, continued. Go back to the study of arsenic contamination in high-arsenic areas in Exercise 14.8.

a. We seek evidence *against* the claim that $\mu=3.2$.$\mu = 3.2.$ What is the sampling distribution of the mean blood arsenic concentration $\bar{x}$ in many samples of 25 adults if the claim is true? Sketch the density curve of this distribution. (Sketch a Normal curve, then mark the axis using what you know about locating the mean and standard deviation on a Normal curve.)

b. Suppose that we take two SRSs of 25 adults residing in two different high-arsenic areas. The first SRS gives $\bar{x} = 3.35\ \mu g/dl$. Use the *P-Value of a Test of Significance* applet to obtain the *P*-value for this area. What would we conclude?

c. The second SRS gives $\bar{x} = 3.75\ \mu g/dl$. What would be the *P*-value for this area? What would we conclude?

d. Explain briefly why these *P*-values tell us that one outcome is strong evidence against the null hypothesis and that the other outcome is not.

14.11 Brains at work. When our brains store information, complicated chemical changes take place. In trying to understand these changes, researchers blocked some processes in brain cells taken from rats and compared these cells with a control group of normal cells. They report that “no difference was seen” between the two groups ($P = 0.45$).[5]

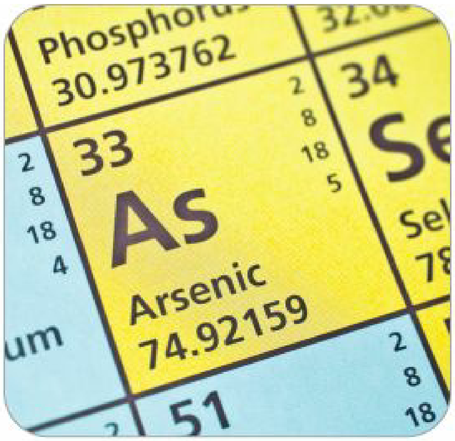
a. State clearly what the *P*-value $P = 0.45$ says about the response that was observed.

b. It isn’t literally true that “no difference was seen.” That is, the computed mean response was not exactly the same in the two groups. Explain what the researchers mean when they give $P = 0.45$ and say “no difference was seen.”

14.12 Arsenic contamination, continued. Go back to Exercise 14.10.

a. Would the outcome for the first SRS be statistically significant at the $\alpha = 0.05$ level? At the $\alpha = 0.01$ level?

b. Would the outcome for the second SRS be statistically significant at the $\alpha = 0.05$ level? At the $\alpha = 0.01$ level?


Phosphorus
30.973762
32.06
34
Se
Sel
78
2
8
18
5
33
As
Arsenic
74.92159
2
8
18
4
um
51
2
8
18
2

William Andrew/Getty Images

TESTS FOR A POPULATION MEAN

We have used tests for hypotheses about the mean μ of a population, under the “simple conditions,” to introduce statistical tests. The test compares the sample mean $\bar{x}$ with the claimed population mean stated by the null hypothesis H_0. The *P*-value says how unlikely an $\bar{x}$ this extreme would be if H_0 was true. Now we can reduce this test to a rule.

z TEST FOR A POPULATION MEAN

Draw an SRS of size *n* from a Normal population that has unknown mean μ and known standard deviation σ. A test statistic and a *P*-value are obtained to **test the null hypothesis that μ has a specified value**

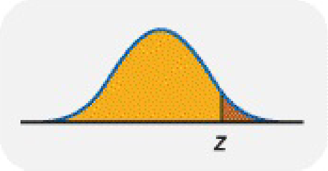
$$H_0 : \mu = \mu_0$$

The **one-sample *z* test statistic** is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

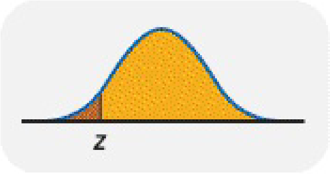
As the variable *Z* follows the standard Normal distribution, the *P*-value for a test of H_0 against

$H_a : \mu > \mu_0$ is $P(Z \geq z)$



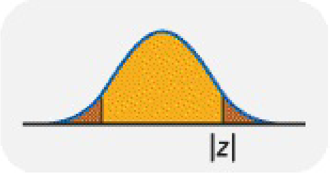
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

$H_a: \mu < \mu_0$ is $P(Z \leq z)$



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

$H_a: \mu \neq \mu_0$ is $2P(Z \geq |z|)$



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Any technology with basic statistical inference capabilities will perform all necessary computations and produce both *z* and the *P*-value.

The test statistic z measures how far the observed sample mean $\bar{x}$ deviates from the hypothesized population value μ_0, in relative terms. This relative measurement is in the standard scale obtained by dividing by the standard deviation of $\bar{x}$. When H_0 is true, z has the standard Normal distribution (which we first described in Chapter 11). The pictures that illustrate the *P*-value look just like Figures 14.6 and 14.7 except that they are in the standard scale.

Many statistical software packages call a *P*-value obtained for a one-sided H_a a "one-sided" or **"one-tailed"** *P*-value. A *P*-value obtained for a two-sided H_a is then called a "two-sided" or **"two-tailed"** *P*-value. The logic for this naming convention is apparent in Figures 14.6 and 14.7: A *P*-value computed for a two-sided H_a includes both "tails" of the Normal sampling distribution, whereas a *P*-value computed for a one-sided H_a includes only one of these "tails."

The steps for carrying out a hypothesis test mirror the overall four-step process for organizing realistic statistical problems.

HYPOTHESIS TESTS: THE FOUR-STEP PROCESS

STATE: What is the practical question that requires a statistical test?

PLAN: Identify the parameter, state the null and alternative hypotheses, and choose the type of test that fits your situation.

SOLVE: Carry out the test in two phases:

1. **Check the conditions** for the test you plan to use.
2. Obtain the **test statistic** and the ***P*-value**.

CONCLUDE: Return to the practical question to describe your results in this setting.

If you do not have access to technology capable of performing inference procedures, sketch the Normal sampling distribution and the area or areas under it representing the *P*-value. Then use an applet for Normal computations or a printed table such as Table B or Table C at the back of the book to obtain the test *P*-value (keep in mind that printed tables can offer only approximate *P*-values).

EXAMPLE 14.9 Body temperature

STATE: When you search the internet for what constitutes “normal” or healthy body temperature, you will probably find that an oral temperature of 98.6 degrees Fahrenheit (°F), or 37.0 degrees Celsius, is considered normal. This widely quoted value comes from a paper published in 1868 by German physician Carl Wunderlich, in which he reported more than a million body temperature readings. In this paper, Wunderlich stated that the mean body temperature of healthy adults is $98.6\ ^{\circ}$ F.$98.6\,^{\circ}\mathrm{F}$. More than a century later, a study was designed to evaluate this claim. It found a mean oral temperature of x¯=98.25 $\bar{x}=98.25\ ^{\circ}$ F$^{\circ}\mathrm{F}$ from 130 adults.[6] Does this study provide significant evidence that Wunderlich’s claim of a mean adult body temperature of $98.6\ ^{\circ}$ F$98.6\,^{\circ}\mathrm{F}$ is not correct?

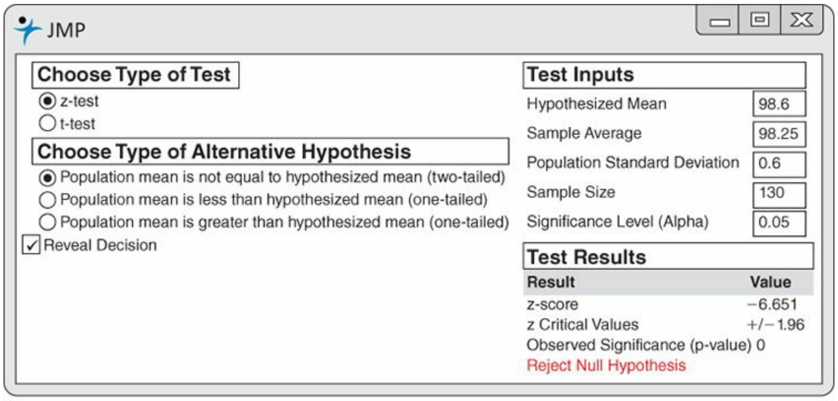
PLAN: The null hypothesis is “no difference” from the accepted mean $\mu_0 = 98.6\,^\circ\mathrm{F}$. The alternative is two-sided because the study did not have a particular direction in mind before examining the data. Thus, the hypotheses about the unknown mean μ of the population of healthy adults are

$$\begin{aligned} H_0 &: \mu = 98.6 \\ H_a &: \mu \neq 98.6 \end{aligned}$$

SOLVE: Check the conditions for inference. Here are the “simple conditions” stated on page 348.

1. **SRS:** The most important condition is that the 130 healthy adults in the sample are an SRS from the population of all healthy adults. We should check this requirement by asking how the data were obtained. The researchers measured the baseline oral temperature of volunteers participating in various vaccine clinical trials conducted at the University of Maryland. If these individuals enrolled in the clinical trial because of a specific medical concern, for example, the data may not be representative of the whole healthy-adult population and could be biased. The researchers assure us that the volunteers were all healthy. Because the subjects were volunteers, the sample is a random sample but not a true SRS. This is very common in the life sciences, especially when dealing with human subjects.
2. **Normal distribution:** We should examine the distribution of the 130 observations to look for signs that the population distribution is not Normal. A histogram in the published report indicates that the data were roughly Normal.
3. **Known σ:** It is unrealistic to suppose that we know that $\sigma = 0.6\,^\circ\mathrm{F}$. This value is based on Wunderlich’s original report of a very large number of body temperatures.

SOLVE: Obtain the test statistic and *P*-value. Figure 14.8 shows the output for a one-sample *z* test using statistical software JMP. It gives $z = -6.651$ and a *P*-value rounded to zero, based on which the output specifies that we should “reject the null hypothesis.”



JMP

Choose Type of Test

● z-test

○ t-test

Choose Type of Alternative Hypothesis

● Population mean is not equal to hypothesized mean (two-tailed)

○ Population mean is less than hypothesized mean (one-tailed)

○ Population mean is greater than hypothesized mean (one-tailed)

☑ Reveal Decision

Test Inputs

Hypothesized Mean	98.6
Sample Average	98.25
Population Standard Deviation	0.6
Sample Size	130
Significance Level (Alpha)	0.05

Test Results

Result	Value
z-score	−6.651
z Critical Values	+/−1.96
Observed Significance (p-value)	0

Reject Null Hypothesis

Figure 14.8

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 14.8 Software output (JMP) for the two-sided test in Example 14.9. The observed value of the test statistics is $z = -6.65$, 6.65, and the two-sided *P*-value is roughly zero.

CONCLUDE: If μ really was $98.6\,^\circ\mathrm{F}$, it would be nearly impossible to obtain an SRS of size 130 from the healthy adult population with sample mean at least as far from $98.6\,^\circ\mathrm{F}$ as that obtained here. The observed $\bar{x} = 98.25\,^\circ\mathrm{F}$ is therefore very strong evidence that the true mean body temperature of healthy adults is not $98.6\,^\circ\mathrm{F}$.

The study findings are highly statistically significant ($P \approx 0$). This *doesn’t* imply that the true mean body temperature in the population of healthy adults is very different from $98.6\,^\circ\mathrm{F}$. Rather, it just says that we have very strong evidence to reject H_0. Because the sample size is rather large ($n = 130$), the sampling distribution of $\bar{x}$ is quite narrow. The observed sample mean $\bar{x} = 98.25\,^\circ\mathrm{F}$ is therefore much further away from a hypothetical population mean of $98.6\,^\circ\mathrm{F}$ than would be expected under such a narrow sampling distribution.

The computations performed by any statistical software in this example involve the following steps:

- Test statistic: $z=(\bar{x}-\mu_0)/(\sigma/n)=(98.25-98.6)/(0.6/130)=-6.65$

$$z = (\overline{x} - \mu_0) / (\sigma/\sqrt{n}) = (98.25 - 98.6) / (0.6/\sqrt{130}) = -6.65$$

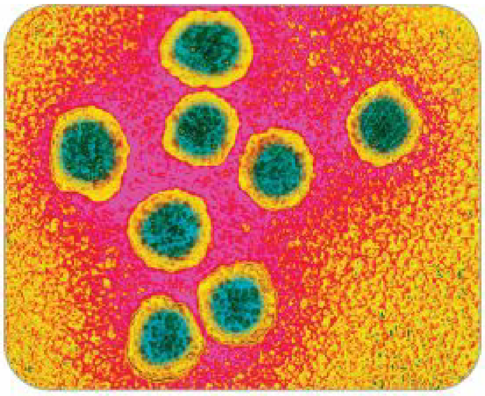
- *P*-value: P=2P(Z≤-6.65)≈0,$P = 2P(Z \leq -6.65) \approx 0,$ computed using the standard Normal curve

This introductory chapter describes the computations involved in statistical inference when the “simple conditions” are met. As we have noted here, these “simple conditions” are not easily met. In Chapter 15 we will discuss circumstances when some of the conditions can be relaxed. In Chapter 17 we will describe a workable alternative to the problem of not knowing the true value of the population standard deviation σ.$\sigma.$

APPLY YOUR KNOWLEDGE

14.13 IQ test scores. Example 14.3 gives the IQ test scores of 31 seventh-grade girls in a Midwest school district. IQ scores follow a Normal distribution with standard deviation σ=15.$\sigma = 15.$ Treat these 31 girls as an SRS of all seventh-grade girls in this district. IQ scores in a broad population are supposed to have mean μ=100.$\mu = 100.$ Is there evidence that the mean in this district differs from 100? Follow the four-step process, as illustrated in Example 14.9, in your answer.

14.14 Hepatitis C and bone density. Does chronic hepatitis C infection affect bone mineral density? In the general population of healthy young adults, bone mineral density is Normally distributed, and density scores are standardized to have a mean μμ of 0 and a standard deviation σσ of 1. A study examined the bone densities of a random sample of 28 adult men living with chronic hepatitis C infection.[7] Their mean standardized bone density was -0.61.$-0.61.$ Does this provide good evidence that the mean standardized bone density of adult men with chronic hepatitis C infection is less than 0? Follow the four-step process as illustrated in Example 14.9.



BSIP/UIG/Getty Images

TESTS FROM CONFIDENCE INTERVALS

Both tests and confidence intervals for a population mean μ start by using the sample mean $\bar{x}$ to estimate μ. Both rely on probabilities calculated from a Normal sampling distribution. In fact, a two-sided test at significance level α can be carried out from a confidence interval with confidence level $C = 1 - \alpha$.

CONFIDENCE INTERVALS AND TWO-SIDED TESTS

A level α two-sided hypothesis test rejects a hypothesis $H_0: \mu = \mu_0$ exactly when the value μ_0 falls outside a level $1 - \alpha$ confidence interval for μ.

EXAMPLE 14.10 Tests from a confidence interval

In Example 14.9, a study found that the mean body temperature for a random sample of 130 healthy adults was $\bar{x} = 98.25$ °F. Is this value significantly different from the textbook value $\mu_0 = 98.6$ °F at the 5% significance level?

We can answer this question directly by performing a two-sided test or indirectly by obtaining a 95% confidence interval. Software gives the interval (98.15, 98.35). The hypothesized value $\mu_0 = 98.6$ falls *outside* this confidence interval, so we *can* reject $H_0: \mu = 98.6$ at the 5% significance level.

Let's examine a different situation, purely for the sake of argument: If the population mean temperature quoted in textbooks had been 98.3 °F, the value tested in H_0 would have fallen *inside* the confidence interval obtained with these data, and we would not have been able to reject the claim $H_0: \mu = 98.3$ at the 5% level.

Statisticians argue over the relative merits of computing a confidence interval or a *P*-value. A hypothesis test is obviously particularly well suited when you have a clearly defined set of hypotheses. The *P*-value will say how strong the evidence against the null hypothesis is. However, if you reject H0,H_0, you are left not knowing what kind of value the parameter might have. And if you fail to reject H0,H_0, there are still other likely values for the parameter besides that defined by H0.H_0. That is, the *P*-value gives no information about effect size (we will discuss this issue further in Chapter 14). *All things being equal, a confidence interval provides more useful information than a P-value alone.*

APPLY YOUR KNOWLEDGE

14.15 Test and confidence interval. The *P*-value for a two-sided test of the null hypothesis H0:μ=10$H_0 : \mu = 10$ is 0.06.

a. Does the 95% confidence interval include the value 10? Why?
b. Does the 90% confidence interval include the value 10? Why?

14.16 Body temperature. Examples 14.9 and 14.10 computed a *P*-value and a confidence interval from data about body temperatures of healthy adults. What can be concluded from each method? What can be concluded from the confidence interval but not from the hypothesis test? What can be concluded from the hypothesis test but not from the confidence interval?

14.17 Confidence interval and test. A 95% confidence interval for a population mean is 31.5±3.5.31.5 ± 3.5.

a. With a two-sided alternative, can you reject the null hypothesis that μ=34$\mu = 34$ at the 5% significance level? Why?
b. With a two-sided alternative, can you reject the null hypothesis that μ=36$\mu = 36$ at the 5% significance level? Why?

CHAPTER 14 SUMMARY

- A **confidence interval** uses sample data to estimate an unknown population parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.
- Any confidence interval has two parts: an interval calculated from the data and a confidence level *C*. The **interval** often has the form

  $$\text{estimate} \pm \text{margin of error}$$

  $\text{estimate} \pm \text{margin of error}$

- The **confidence level** is the success rate of the method that produces the interval. That is, *C* is the probability that the method will give a correct answer. If you use 95% confidence intervals often, in the long run 95% of your intervals will contain the true value of the parameter. You do not know whether a 95% confidence interval calculated from a particular set of data contains the true parameter value.
- A level *C* **confidence interval for the mean** μ of a Normal population with known standard deviation σ, based on an SRS of size n, is given by

  $$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

  $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$

  The **critical value** z^* is chosen so that the standard Normal curve has area *C* between $-z^*$ and z^*. Confidence intervals are typically obtained using technology (applet or software), although the computations can also be worked out from the equation and the appropriate critical value.
- A **hypothesis test** assesses the evidence provided by data against a **null hypothesis** H_0 and in favor of an **alternative hypothesis** H_a.
- Hypotheses are always stated in terms of population parameters. Usually, H_0 is a statement that no effect is present,[8] and H_a says that a parameter differs from its null value in a specific direction (**one-sided alternative**) or in either direction (**two-sided alternative**).

- The essential reasoning of a hypothesis test is as follows. Suppose for the sake of argument that the null hypothesis is true. If we repeated our data production many times, would we often get data as inconsistent with H_0 as the data we actually have? Data that would be unlikely if H_0 was true provide evidence against H_0.
- A hypothesis test is based on a **test statistic** that measures how far (relative to expected variability) the sample outcome is from the value stated by H_0.
- The ***P*-value** of a test is the probability, computed supposing H_0 to be true, that the test statistic would take a value at least as extreme as that actually observed. Small *P*-values indicate strong evidence against H_0. The computations for the *P*-value rely on the sampling distribution of the test statistic when H_0 is true.
- If the *P*-value is as small as or smaller than a specified value α, we say that the data are **statistically significant** at significance level α.
- **A hypothesis test for the null hypothesis** $H_0: \mu = \mu_0$ concerning the unknown mean μ of a population is based on the **one-sample *z* test statistic**

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

  The *z* test assumes an SRS of size n from a Normal population with known population standard deviation σ. The test *P*-value is typically obtained using technology (applet or software), although the computations can also be worked out with the preceding equation and the use of the standard Normal distribution.

THIS CHAPTER IN CONTEXT

The reason we collect data is not to learn about the individuals that we examined, but rather to make infererences from the data about some wider population that the individuals represent. In this chapter we discuss the basic reasoning of statistical inference—confidence intervals and hypothesis tests—with emphasis on statistical inference for a population mean under simple conditions.

The computations described in this chapter build on the Normal calculations you learned in Chapter 11 and on the properties of the sampling distribution of $\bar{x}$ introduced in Chapter 13. We examine other methods for computing confidence intervals in Chapters 17 through 20 and Chapter 26. All inference chapters (Chapters 17 to 28) in this book describe some specific computational method for testing hypotheses and obtaining *P*-values. While the mathematical approach varies across these chapters, the interpretation of confidence intervals and hypothesis tests remains fundamentally the same.

In the next chapter we discuss some limitations to obtaining and interpreting confidence intervals and hypothesis tests. As with Chapter 14, the essence of Chapter 15 is directly applicable to all the inference chapters in the remainder of the book. In the optional sections of Chapter 15, we also consider some more mathematically sophisticated approaches relevant to study planning.

CHECK YOUR SKILLS

14.18 The Framingham Heart Study recorded the systolic blood pressure (SBP) of 3,534 adults and reported a 95% confidence interval for the population mean SBP of 126.7 to 127.9 mm Hg. The correct interpretation for this confidence interval is that we are 95% confident that

a. the adults in this sample have SBP values between 126.7 and 127.9 mm Hg.
b. adults in the population have SBP values between 126.7 and 127.9 mm Hg.
c. the population mean SBP is a value between 126.7 and 127.9 mm Hg.

14.19 What is the margin of error for the cited confidence interval of 126.7 to 127.9 mm Hg?

a. 0.6 mm Hg
b. 1.2 mm Hg
c. 2.4 mm Hg

14.20 A study reports the mean change in high-density lipoprotein (HDL, also known as “good” cholesterol) of adults eating raw garlic six days a week for six months. The margin of error for a 95% confidence interval is given as plus or minus 6 milligrams per deciliter of blood (mg/dl).[9] This means that

a. we can be certain that the study result is within 6 mg/dl of the true mean HDL change in this population.
b. we can be certain that the study result is within 6 mg/dl of the true mean HDL change in this population if the conditions for inferences are satisfied.
c. the study used a method that gives a result within 6 mg/dl of the true mean HDL change in this population for 95% of all samples of 3,534 adults.

Use the following information for Exercises 14.21 and 14.22. A laboratory scale is known to have a standard deviation of $\sigma=0.005$ gram$\sigma = 0.005\ gram$

in repeated weighings. Scale readings in repeated weighings are Normally distributed, with the mean of all possible measurements equal to the true weight of the specimen.

14.21 A specimen is weighed 4 times on this scale. The average weight is 4.53 grams. What is a 99% confidence interval for the true weight of this specimen?

a. (4.521, 4.539)
b. (4.524, 4.536)
c. (4.527, 4.533)

14.22 Suppose that we needed a 95% confidence level for this specimen using the same data. What would the margin of error be?

a. 0.0025 grams
b. 0.0049 grams
c. 0.0069 grams

Use the following information for Exercises 14.23 through 14.25. The average human gestation time is 266 days from conception. A researcher suspects that proper nutrition plays an important role and that poor women with inadequate food intake would have shorter gestation times even when given vitamin supplements. A random sample of 20 poor women given vitamin supplements throughout their pregnancy has a mean gestation time from conception of $\bar{x}=256$ days.$\bar{x} = 256\,days.$

14.23 What is the null hypothesis for the researcher's test?

a. $H0:\mu=266$$H_0 : \mu = 266$
b. $H0:\mu=256$$H_0 : \mu = 256$
c. $H0:\mu<266$$H_0 : \mu < 266$

14.24 What is the researcher's alternative hypothesis for this test?

a. $Ha:\mu\neq256$$H_a : \mu \neq 256$
b. $Ha:\mu<266$$H_a : \mu < 266$
c. $Ha:\mu<256$$H_a : \mu < 256$

14.25 Human gestation times are approximately Normal with standard deviation $\sigma=16$ days.$\sigma = 16\,\text{days}.$ The *P*-value for the researcher's test is

a. more than 0.1.
b. between 0.01 and 0.1.
c. less than 0.01.

14.26 You use software to run a hypothesis test. The program tells you that the *P*-value is 0.031.0.031. This result is

a. not significant at the 5% level.
b. significant at the 5% level but not at the 1% level.
c. significant at the 1% level.

14.27 Average human gestation time is 266 days, when counted from conception. A hospital gives a 90% confidence interval for the mean gestation time from conception among its patients: 264±5$264 \pm 5$ days. Is the mean gestation time in that hospital significantly different from 266 days?

a. It is not significantly different at the 10% level and, therefore, is also not significantly different at the 5% level.
b. It is not significantly different at the 10% level but might be significantly different at the 5% level.
c. It is significantly different at the 10% level.

CHAPTER 14 EXERCISES

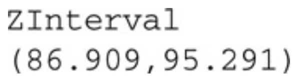
14.28 Explaining confidence. You calculate a 95% confidence interval of 27 ± 2 centimeters (cm) for the mean needle length of Torrey pine trees. You ask friends to explain this result.

a. One friend believes it means that "95% of all Torrey pine needles have lengths between 25 and 29 cm." Is that right? Explain your answer.
b. Another friend thinks it means that "we can be 95% confident that the true mean needle length of Torrey pine trees is 27 cm." Is that right? Explain your answer.

14.29 Explaining confidence. Here is an explanation from the Associated Press concerning one of its opinion polls. Explain briefly but clearly in what way this explanation is incorrect.

> *For a poll of 1,600 adults, the variation due to sampling error is no more than three percentage points either way. The error margin is said to be valid at the 95 percent confidence level. This means that, if the same questions were repeated in 20 polls, the results of at least 19 surveys would be within three percentage points of the results of this survey.*

14.30 Deer mice. Deer mice (*Peromyscus maniculatus*) are small rodents native to North America. Their adult body lengths (excluding tail) are known to vary approximately Normally, with mean $\mu = 86$ millimeters (mm) and standard deviation $\sigma = 8$ mm.[10] Deer mice are found in diverse habitats and exhibit different adaptations to their environment. A random sample of 14 deer mice in a rich forest habitat gives an average body length of $\bar{x} = 91.1$ mm. Assume that the standard deviation σ of all deer mice in this area is also 8 mm. Here is the output from the TI-83 graphing calculator for a 95% confidence interval based on these findings:



```
ZInterval
(86.909,95.291)
```

a. What is captured by this confidence interval?
b. Express this confidence interval as a sentence written in the context of this problem.
c. What is the margin of error for this confidence interval? Express this interval in the format “estimate±margin of error.estimate ± margin of error.”

14.31 Deer mice, continued. Exercise 14.30 gives a 95% confidence interval for the mean body length of deer mice living in a rich forest habitat.

a. What is the standard deviation of the sampling distribution of mean body lengths $\bar{x}$?$\bar{x}$?
b. What critical value was used to compute this 95% confidence interval?
c. Show the step-by-step computations required to arrive at the interval provided by the graphing calculator.
d. Would a 90% confidence interval based on the same data be larger or smaller? Explain your reasoning.

14.32 Waist circumference. The 2012 National Health and Nutrition Examination Survey reports a 95% confidence interval of 99.8 to 102.0 centimeters for the mean waist circumference of adult women in the United States.

a. What is captured by this confidence interval?
b. Express this confidence interval as a sentence written in the context of this problem.
c. What is the margin of error for this confidence interval? Express this interval in the format “estimate±margin of error.estimate ± margin of error.”
d. Would a 99% confidence interval based on the same data be larger or smaller? Explain your reasoning.

14.33 Avandia and cholesterol. Drug maker GlaxoSmithKline investigated the potential impact of its oral antidiabetic drug Avandia (rosiglitazone maleate) on the blood lipid levels of adults diagnosed with diabetes who might benefit from taking Avandia. The mean baseline low-density lipoprotein (LDL, also known as “bad” cholesterol) level for a random sample of 964 such patients was $\bar{x}=125.6$ milligrams per deciliter (mg/dl).$\bar{x} = 125.6$ milligrams per deciliter (mg/dl). The distribution of LDL levels in the adult population is known to be close to Normal with standard deviation σ=30 mg/dl.$\sigma = 30$ mg/dl.[11] Here is a 99% confidence interval for the mean baseline LDL level μμ among adult diabetics who might benefit from taking Avandia, assuming that σσ is the same in this population as in the general population.

```
One-Sample Z
The assumed standard deviation = 30
  N      Mean  SE Mean          99% CI
964  125.600    0.966  (123.111, 128.089)
```

a. Express this confidence interval as a sentence written in the context of this problem.

b. One of the “simple conditions” for inference about a population mean is that the population standard deviation σσ is known. Explain why this assumption may not be reasonable in this specific example.

14.34 Student heights. In Example 11.1 (page 274) we saw that the distribution of heights of young women aged 18 to 24 is approximately Normal with mean μ=64.5 inches$\mu = 64.5$ inches and standard deviation σ=2.5 inches.$\sigma = 2.5$ inches. A random sample of 93 female undergraduate students at the University of California at Irvine produced a sample mean height of 62.7 inches. The individual heights are displayed in Table 14.1.[12]

TABLE 14.1 Heights in inches of female undergraduate students at University of California at Irvine

54	54	56	57	58	58	58	59	60	60	60	60	60	60	60	60	61	61	61
61	61	61	61	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62
63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63	63
63	63	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	64	65
65	65	65	65	65	65	65	66	66	66	67	67	67	67	68	68	70		

a. Assume that the standard deviation of young women’ heights is the same at UC–Irvine as in the rest of the United States (σ=2.5 inches).($\sigma = 2.5$ inches). Briefly discuss the other two simple conditions, using a histogram or a dotplot to verify that the distribution is roughly symmetric with no outliers.

b. Following the four-step process, give a 95% confidence interval for the mean height of all female undergraduate students at UC–Irvine.

14.35 Sodium intake. The Institute of Medicine sets the tolerable upper level of sodium intake for adults and teenagers at 2300 milligrams (mg) per day. A random sample of 3727 adolescents representative of the U.S. adolescent population found a mean sodium intake of 3486 mg per day. Does this finding suggest that U.S. adolescents consume sodium in excess of the tolerable upper level? State H0H_0 and HaH_a for a test to answer this question. State carefully what the parameter μμ in your hypotheses is.

14.36 Cartons of eggs. According to the U.S. Department of Agriculture, cartons of 12 large eggs should have a minimum weight of 24 ounces (oz), which is equivalent to an average egg weight of 2 oz or more.[13] You suspect that a certain producer sells eggs of inferior caliber (lower weight) in its cartons labeled “large eggs.” State H0H_0 and HaH_a for a test to assess your suspicion. State what the parameter μμ in your hypotheses is.

14.37 The wrong alternative. One of your friends is testing the effect of drinking coffee on the duration of cold symptoms. The common cold lasts, on average, 6 days. Your friend starts with no expectations as to whether drinking coffee will have any effect on cold duration. After seeing the results of the experiment, in which the average cold duration was less than 6 days, your friend tests a one-sided alternative about the population mean cold duration when drinking coffee,

H0:μcoffee=6Ha:μcoffee<6

$$\begin{aligned}&\mathrm{H}_0 : \mu_{coffee} = 6\\&\mathrm{H}_a : \mu_{coffee} < 6\end{aligned}$$

which gives z=−1.68$z = -1.68$ with one-sided *P*-value P=0.0465.$P = 0.0465.$

a. Explain why your friend should have used a two-sided alternative hypothesis.

b. What is the correct two-sided *P*-value for z=-1.68?$z = -1.68?$

14.38 Hydration. Researchers used data from the National Health and Nutrition Examination Survey to evaluate the hydration status of American youths (children and adolescents age 6 to 19 years).[14] Adequate hydration corresponds to a urine osmolality of less than 800 milliosmoles per kilogram (mOsm/kg). The researchers wanted to know if, on average, American youths are adequately hydrated. The study's nationally representative sample of 4134 youths had a mean urine osmolality of 755.5 mOsm/kg. State the corresponding null and alternative hypotheses, using appropriate math symbols.

14.39 Is this what significance means? Statistical significance is very often mentioned in research reports. When asked what it means, a student replies that "results that are statistically significant tell us that they cannot easily be explained by chance variation alone if the hypothesis tested was true." Do you think that this statement is essentially correct? Explain your answer.

14.40 Is this what *P* means? When asked to explain the meaning of "the *P*-value was 0.03,$0.03,$" a student says, "It means there is only probability 0.03 that the null hypothesis is true." Is this an essentially correct explanation? Explain your answer.

14.41 Significance level: 5% versus 1%. Sketch the standard Normal curve for the *z* test statistic and mark off areas under the curve to show why a value of *z* that is significant at the 1% level in a one-sided test is always significant at the 5% level. If *z* is significant at the 5% level, what can you say about its significance at the 1% level?

14.42 Gender-orientation of smells. Some smells are perceived as being more feminine or more masculine. Researchers asked a random sample of 300 American adults to score quantitatively the perceived gender-orientation of lavender. Smaller scores indicate a more feminine scent and larger scores a more masculine scent. In this design, a gender-neutral scent would have a perceived gender-orientation score of 12. The participants gave the scent of lavender an average score of 10.28.[15]

a. Is this evidence that the scent of lavender is not gender-neutral, on average? State H0H_0 and HaH_a in the corresponding test, indicating what the parameter in your hypotheses is.

b. The researchers report that “p<.01$p < .01$ (2-tail)” for this test. Explain the meaning of “p<.01$p < .01$” and “2-tail.”

c. Assuming that the conditions for inference were met, conclude in the context of this study.

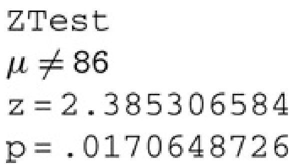
14.43 Gender-orientation of smells, continued. Researchers in the study described in Exercise 14.42 also examined the perceived gender-orientation of cinnamon. The participants gave the scent of cinnamon an average score of 11.76.

a. Is this evidence that the scent of cinnamon is not gender-neutral, on average? State H0H_0 and HaH_a in the corresponding test, indicating what the parameter in your hypotheses is.

b. The researchers report that this test was not statistically significant. Explain why we cannot conclude that cinnamon is a gender-neutral scent.

c. Assuming that the conditions for inference were met, conclude in the context of this study.

14.44 Deer mice, continued. The sample of 14 deer mice in Exercise 14.30 had an average body length of x¯=91.1 mm.$\bar{x} = 91.1\text{ mm}.$ Do forest deer mice on the average differ significantly in body length from the general deer mouse population (which is known to have mean body length μ=86 mm$\mu = 86\text{ mm}$)? Assuming that the standard deviation σσ for all deer mice in this area is also 8 mm, here is the output from the TI-83 graphing calculator for the corresponding two-sided hypothesis test based on these findings:



```
ZTest
μ≠86
z=2.385306584
p=.0170648726
```

a. Explain why the appropriate test in this situation should be two-sided.

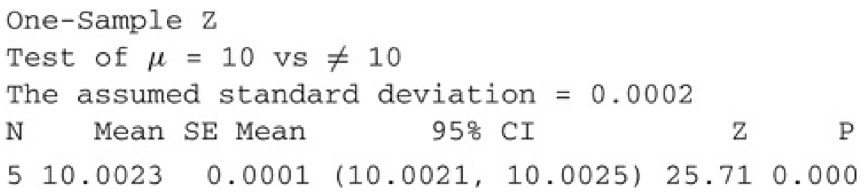
b. What are the values z and p in the calculator output?

c. Assuming that the conditions for inference were met, conclude in the context of this study.

14.45 Deer mice: hypothesis test and confidence interval. Exercise 14.30 gave a 95% confidence interval for the mean body length μμ of deer mice in a rich forest habitat. Exercise 14.44 used the same data to test the hypothesis that deer mice from this habitat have significantly different body lengths from the general population of deer mice. Compare the two results and explain why the confidence interval is more informative than the test *P*-value.

14.46 Student heights, continued. Exercise 14.34 (along with Table 14.1) provides the heights of a random sample of 93 female undergraduate students at the University of California at Irvine. Assuming that heights in this undergraduate population are Normal with standard deviation σ=2.5 inches,$\sigma = 2.5\text{ inches,}$ do the data provide evidence that the mean height of all female undergraduate students at UC–Irvine is different from the mean height of 64.5 inches of all young women in the United States? Follow the four-step process, as illustrated in Example 14.9.

14.47 Hypothesis tests and confidence intervals. To assess the accuracy of a laboratory scale, a reference weight known to weigh exactly 10 grams (g) is weighed repeatedly. The scale readings are Normally distributed with standard deviation σ=0.0002 g.$\sigma = 0.0002\text{ g.}$ The reference weight is weighed 5 times on that scale. The mean result is 10.0023 g. Here is the output from the statistical software Minitab for these data:



```
One-Sample Z
Test of μ = 10 vs ≠ 10
The assumed standard deviation = 0.0002
N    Mean  SE Mean       95% CI            Z     P
5 10.0023   0.0001 (10.0021, 10.0025) 25.71 0.000
```

a. Do the 5 weighings give good evidence that the scale is not well calibrated (that is, its mean μμ for weighing this weight is not 10 g)? Explain your answer.
b. Give a 95% confidence interval for the mean weight on this scale for all possible measurements of the reference weight. What do you conclude about the calibration of this scale?
c. Compare your conclusions in parts a and b. Explain why the confidence interval is more informative than the test result.

CHAPTER 15 Inference in Practice

Vimeo
Google+
Reddit
Facebook
Instagram
LinkedIn
Twitter
1

Twin Design/Shutterstock.com

IN THIS CHAPTER WE COVER...

- Conditions for inference in practice
- How confidence intervals behave
- How hypothesis tests behave
- *Discussion: The scientific approach*
- Planning studies: selecting an appropriate sample size

To this point, we have learned just two procedures for statistical inference. Both concern inference about the mean μ of a population when the “simple conditions” (page 348) are true: *The data are from an SRS, the population has a Normal distribution, and we know the standard deviation σ of the population*. Under these simple conditions, a confidence interval for the mean μ is $\bar{x} \pm m$, in which the margin of error m uses a critical value z^* as a multiplier. To test a hypothesis $H_0: \mu = \mu_0$, we obtain the one-sample z statistic and corresponding *P*-value. We call these inference procedures ***z* procedures** because they both start with the one-sample z statistic and use the standard Normal distribution. The simple conditions and the reliance on Normal computations make the mathematics of inference simpler and easier to understand. This is why we approached statistical inference under such conditions first.

In later chapters we will modify these procedures for inference about a population mean to make them useful in practice. We will also introduce procedures needed beyond the one-sample situation that are commonly used to analyze either quantitative or categorical data from observational and experimental studies. There are libraries—both of books and of software—full of elaborate statistical techniques that are beyond the scope of an introductory statistics textbook. Yet, the reasoning of confidence intervals and hypothesis tests (introduced in Chapter 14) and the practical issues of statistical inference (covered in this chapter) are fundamentally the same for all inference procedures, no matter how elaborate they get.

Statisticians have a saying: “Mathematical theorems are true; statistical methods are effective when used with judgment.” That the one-sample z statistic has the standard Normal distribution when the null hypothesis is true is a mathematical theorem. Effective use of statistical methods requires more than knowing such facts. This chapter begins the process of helping you develop the judgment needed to use statistics in practice. That process will continue in examples and exercises through the rest of this book.

CONDITIONS FOR INFERENCE IN PRACTICE

Any confidence interval or hypothesis test can be trusted only under specific conditions. It's up to you to understand these conditions and to judge whether they fit your problem. With that caveat in mind, let's look back at the "simple conditions" for the z confidence interval and test for one population mean.

The final "simple condition"—that we know the population standard deviation σ of the quantitative variable studied—is rarely satisfied in practice. The z procedures are therefore of little practical use, especially in the life sciences. Fortunately, it's easy to do without the "known σ" condition; Chapter 17 shows how. The first two "simple conditions" (SRS, Normal population) are more difficult to escape if we are to trust the results of statistical inference. As you plan inference, you should always ask, "Where did the data come from?" For a quantitative variable, you must often also ask, "What is the shape of the population distribution?" This is the point where knowing mathematical facts gives way to the need for judgment.

Where did the data come from? *The most important requirement for any inference procedure is that the data come from a process to which the laws of probability apply*. Inference is most reliable when the data come from a probability sample or a randomized comparative experiment. Probability samples use chance to choose respondents. Randomized comparative experiments use chance to assign subjects to treatments. The deliberate use of chance ensures that the laws of probability apply to the outcomes, which in turn ensures that statistical inference makes sense.

WHERE THE DATA COME FROM MATTERS

When you use statistical inference, you are acting as if your data are a probability sample or come from a randomized experiment.

If your data don't come from a probability sample or a randomized comparative experiment, your conclusions may be challenged. To answer the challenge, you must usually rely on subject-matter knowledge, not on statistics. It is common practice to apply statistics to data that are not produced by true random selection. When you see such a study, ask whether the data can be trusted as a basis for the conclusions of the study.

EXAMPLE 15.1 The neurobiologist and the sociologist

A neurobiologist is interested in how our visual perception can be fooled by optical illusions. Her subjects are students in Neuroscience 101 at her university. Most neurobiologists would agree that it's safe to treat the students as an SRS of all people with normal vision. There is nothing unique about being a student that changes the neurobiology of optical illusions.

A sociologist at the same university uses students in Sociology 101 to examine attitudes toward the use of human subjects in science. Students as a group are younger than the adult population as a whole. Even among young people, students as a group tend to come from more prosperous and better-educated homes. Even among students, this university isn't typical of all campuses. Even on this campus, students in a sociology course may have opinions that are quite different from those of engineering students or biology students. The sociologist can't reasonably act as if these students are a random sample from a population of general interest such as all adults or even all young adults in college.

Our first examples of inference, using the z procedures, assumed that each data set was an SRS from the population of interest. Let's look back at the examples in Chapter 14.

EXAMPLE 15.2 Is it really an SRS?

The NHANES study that produced the height data for Example 14.1

(page 348) uses a complex multistage sample design, so it’s a bit of an oversimplification to treat the height data as coming from an SRS from the population of eight-year-old boys.[1] Overall, the NHANES sample closely resembles an SRS, but professional statisticians would use more complex inference procedures to match the more complex design of the sample.

The 31 seventh-grade girls in the IQ scores study in Example 14.3 were randomly chosen from the population of seventh-grade girls in a Midwest school district. We can treat these girls as an SRS from the population of interest.

The study of inorganic phosphorus in Example 14.4 uses existing medical records from a random sample of 12 elderly men and women between the ages of 75 and 79 years. However, all these medical records came from the archives of a single laboratory in Ohio. This is a major problem for two reasons. First, individuals who have their blood tested in Ohio may not be representative of the whole population of Americans aged 75 to 79 years and of that population’s overall susceptibility to various medical conditions. Second, different laboratories are known to produce slightly different blood test results. These two issues should be clearly discussed when presenting the results of the statistical analysis because they severely limit our ability to generalize the findings beyond the population of Ohio residents aged 75 to 79 years who would get their blood tested at that particular laboratory.

These examples are fairly typical. Only one involves an actual SRS of the true population of interest. One represents a situation in which we assume an SRS might be used for a quick analysis of data from a more complex probability sample. The last example uses an SRS taken from a small portion of the actual population of interest and may not be appropriate to answer the study’s question. *There is no simple rule for deciding when you can act as if a sample is an SRS. Pay attention to these cautions:*

CAUTION

- *Practical problems such as nonresponse in samples or dropouts from an experiment can hinder inference from even a well-designed study.* NHANES has about an 80% response rate. This is much higher than the response rates for opinion polls and most other national surveys, so by realistic standards NHANES data are quite trustworthy. (NHANES uses advanced methods beyond the scope of an introductory course to try to correct for nonresponse.)

- *Different methods are needed for different designs.* The zz procedures aren’t correct for probability samples more complex than an SRS. Later chapters give methods for some other designs, but we won’t discuss inference for really complex designs like that used by NHANES. Always be sure that you (or your statistical consultant) know how to carry out the inference your design calls for.
- *There is no cure for fundamental study design flaws like voluntary response surveys, uncontrolled experiments, or biased samples.* Look back at the bad examples in Chapters 6 and 7 and steel yourself to just ignore data from such studies.

What is the shape of the population distribution? Parametric inference procedures, like the one-sample zz procedures for a population mean, start with a model for the distribution of the random variable in the population. Every aspect of the distribution is specified in the model, except for one parameter. This parameter can then be estimated with a confidence interval or evaluated with a hypothesis test. The assumptions, or conditions, for inference are the model specifications.

Many of the most basic methods of inference are designed for Normal populations. That’s the case for the zz procedures as well as for the more practical procedures for inference about means that we will examine in Chapters 17 and 18. Fortunately, we have a lot of leeway around this condition.

This flexibility arises because the zz procedures and many other procedures designed for Normal distributions rely on *Normality of the sampling distribution of* x-,$\bar{x}$, not Normality of individual observations in the population. The central limit theorem (page 333) tells us that x-$\bar{x}$ is more Normal than individual observations and that x-$\bar{x}$ is more Normal for larger sample sizes. In practice, the zz procedures are reasonably accurate for any roughly symmetric distribution for samples of even moderate size. If the sample is large, the sampling distribution of x-$\bar{x}$ will be close to Normal even if individual measurements are strongly skewed, as Figure

13.5 (page 332) illustrates. Later chapters give practical guidelines on sample sizes for specific inference procedures.

We rarely know the shape of the population distribution. In practice we rely on previous studies and on the analysis of sample data. Sometimes experience suggests that our data are likely to come from a roughly Normal distribution—or not. For example, heights of people of the same sex and similar ages are close to Normal, but weights are not. Always explore your data before undertaking inference. When the data are chosen at random from a population, the shape of the sample data distribution mirrors the shape of the population distribution. Make a dotplot or a histogram of your data and see whether the shape is roughly Normal, or make a Normal quantile plot if you have access to the necessary technology.

Remember that *small samples have a lot of chance variation,* so that Normality is difficult to judge from just a few observations. Also, outliers in a sample can distort the results of inference. *Any inference procedure based on sample statistics that are not resistant to outliers, like the sample mean* x-,$\bar{x}$, *can be strongly influenced by a few extreme observations.* Always check for outliers and deal with them appropriately (correcting typos or excluding outliers that have a legitimate reason for being excluded) before performing the zz procedures or other inference methods based on statistics like x-$\bar{x}$ that are not resistant to outliers. Read the discussion “Dealing with Outliers” in Chapter 2 (page 55) for detailed examples.

When outliers are present or when the data suggest that the population is strongly non-Normal and your sample size is small, consider alternative methods that don’t require Normality and are not sensitive to outliers. Some of these methods appear in Chapter 27 (on the companion website).

APPLY YOUR KNOWLEDGE

15.1 A medical journal takes a poll. The *New England Journal of Medicine* posts its peer-reviewed articles and editorials on its website. An opt-in poll was featured next to an editorial about the regulation of sugar-sweetened beverages. The poll asked, “Do you support government regulation of sugar-sweetened beverages?” Readers just needed to click on a response (yes or no) to become part of the sample. The poll stayed open for several weeks in October 2012. Of the 1290 votes cast, 864 were “yes” responses.[2]

a. Would it be reasonable to calculate from these data a confidence interval for the percent answering “yes” in the American population? Explain your answer.
b. Would it be reasonable to calculate from these data a confidence interval for the percent answering “yes” among online readers of the *New England Journal of Medicine*? Explain your answer.

15.2 Exploratory data analysis. Tree shrews, *Tupaia belangeri,* are small omnivorous mammals that are phylogenetically related to primates. A research team examined paw preference during grasping tasks among 36 tree shrews born and raised in captivity. A “pawedness index” was computed as a unitless variable taking values between −1 and 1, depending on the relative use of the right and left paws, so that an animal always using its left paw would have a value of −1 but an animal using its right and left paws equally would have a value of 0. The individual pawedness index values are displayed in the following table.[3]

−1.00	−1.00	−1.00	−1.00	−1.00	−0.96	−0.95	−0.89	−0.06
0.86	0.90	1.00	1.00	−1.00	−1.00	−1.00	−0.89	−0.25
0.87	0.94	1.00	1.00	−0.93	−0.03	−1.00	−1.00	−0.08
0.13	0.56	0.67	−0.11	0.24	0.50	1.00	1.00	0.41

Avalon/Bruce Coleman Inc/Alamy Stock Photo

Display these data in a dotplot and describe the distribution of pawedness index values. Explain why a confidence interval for the population mean pawedness index of tree shrews raised in captivity would not be very informative and could easily be misinterpreted. (Note: It is always a good idea to look carefully at your data before obtaining summary statistics and performing statistical inference procedures.)

15.3 Sampling medical records. A retrospective study reported on complications from prostate needle biopsies by examining the medical records of 1000 consecutive cases of patients who underwent the procedure at one academic medical center between September 2001 and August 2010.[4] Why is it risky to regard these 1000 patients as an SRS from the population of all patients who have a prostate needle biopsy? Name some factors that might make these 1000 consecutive patient records unrepresentative of all prostate needle biopsy outcomes.

HOW CONFIDENCE INTERVALS BEHAVE

The confidence interval $\bar{x} \pm m$ for a population mean μ is fundamentally similar to confidence intervals for other parameters that take the form *estimate* $\pm m$, where the margin of error m is obtained by multiplying a measure of spread of the sampling distribution of the estimate by some critical value that corresponds to the chosen confidence level.

Ideally, we would like both high confidence and a small margin of error. High confidence says that our method almost always gives correct answers. A small margin of error says that we have pinned down the parameter quite precisely. The factors that influence the margin of error of the z confidence interval are typical of most confidence intervals.

How do we get a small margin of error? The margin of error for the z confidence interval is

$$\text{margin of error} = z^* \frac{\sigma}{\sqrt{n}}$$

This expression has z^* and σ in the numerator and $\sqrt{n}$ in the denominator. Therefore, the margin of error gets smaller when

CAUTION

- z^* is smaller. A smaller critical value z^* is the same as a lower confidence level C (look at Figure 14.3, page 353). *There is a trade-off between the confidence level and the margin of error. To obtain a smaller margin of error from the same data, you must be willing to accept lower confidence.* This is not useful.

- σ is smaller. The standard deviation σ measures the variation in the population. You can think of the variation among individuals in the population as noise that obscures the average value μ. It is easier to pin down μ when σ is small. Some of that noise may come from measurement errors, so it could be reduced to some extent by using more precise measuring instruments (for example, measuring heights with a precision of 1 millimeter rather than 1 centimeter). A large part of the variation among individuals in a population, however, is simply natural variation that cannot be reduced (for example, individual heights).

- nn is larger. Increasing the sample size nn reduces the margin of error for any confidence level. As a consequence, larger samples allow more precise estimates. However, *because n appears under a square root sign, we must take four times as many observations to cut the margin of error in half*. Practical issues (time, resources, cost, or even ethics) may also limit just how much data can be collected.

EXAMPLE 15.3 Confidence level and margin of error

In Example 14.3 (page 354), we examined the IQ scores of 31 seventh-grade girls in a Midwest school district. The data gave x-=105.84, $\bar{x} = 105.84,$ and we know that σ=15.$\sigma = 15.$ A 95% confidence interval for the mean IQ score for all seventh-grade girls in this district is

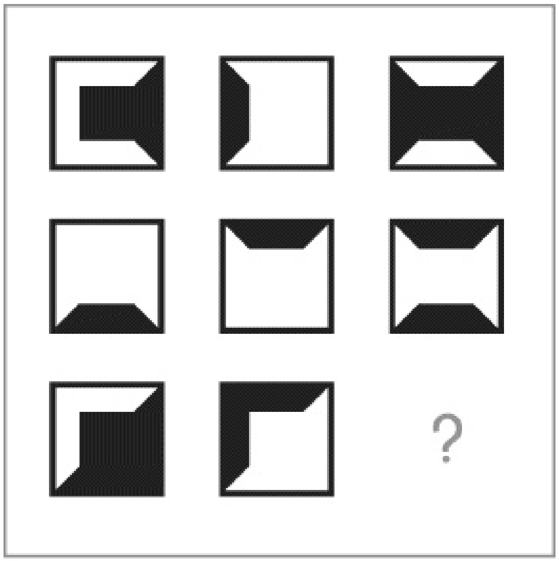
x-±z*σn=105.84±1.9601531=105.84±5.28

$$\begin{aligned} \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= 105.84 \pm 1.960 \frac{15}{\sqrt{31}} \\ &= 105.84 \pm 5.28 \end{aligned}$$

A 90% confidence interval based on the same data replaces the 95% critical value z*=1.960$z^* = 1.960$ with the 90% critical value z*=1.645.$z^* = 1.645.$ This interval is

x-±z*σn=105.84±1.6451531=105.84±4.43

$$\begin{aligned} \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} &= 105.84 \pm 1.645 \frac{15}{\sqrt{31}} \\ &= 105.84 \pm 4.43 \end{aligned}$$



Lyudmyla Kharlamova/Shutterstock.com

Lower confidence results in a smaller margin of error, ±4.43± 4.43 instead of ±5.28.$\pm 5.28.$ In the same way, you can calculate that the margin of error for 99% confidence is larger (±6.94)(± 6.94) because the critical value multiplier is larger (z*=2.576).$(z^* = 2.576).$ Figure 15.1 compares these three confidence intervals.

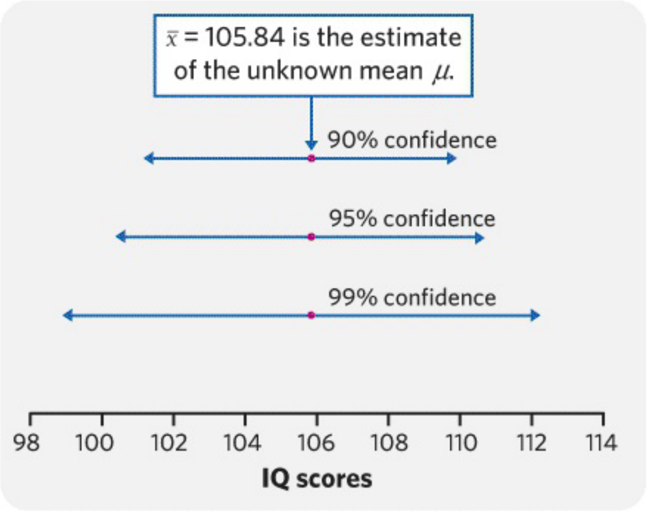


Figure 15.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 15.1 The lengths of three confidence intervals for the IQ study. All three are centered at the estimate x-=105.84.

> $$\bar{x} = 105.84.$$
>
> When the data and the sample size remain the same, higher confidence requires a larger margin of error.

If we had a sample of 62 girls, the margin of error for 95% confidence would decrease from ±5.28 ± 5.28 (1.960×15/31) $(1.960 \times 15/\sqrt{31})$ to ±3.73 ± 3.73 (1.960×15/62). $(1.960 \times 15/\sqrt{62})$. Doubling the sample size does *not* cut the margin of error in half, because the sample size n n appears under a square root sign. Also keep in mind that a sample of a different size would most likely produce a different sample mean, so that the center of the confidence interval would be affected as well.

Throughout this book, we will see several procedures used to obtain confidence intervals for various parameters under different data collection designs. Regardless of the procedure, all these confidence intervals are affected by the chosen confidence level, data variability, and sample size.

The margin of error accounts only for sampling error The most important caution about confidence intervals in general is a consequence of the use of a sampling distribution. A sampling distribution describes how a statistic such as x- $\bar{x}$ varies in repeated random samples. This variation causes “random sampling error” because the statistic misses the true parameter by a random amount. Sampling distributions do not take into account any other source of variation or bias in the sample data. Thus, *the margin of error in a confidence interval ignores everything except the sample-to-sample variation due exclusively to choosing the sample randomly.*

THE MARGIN OF ERROR DOESN’T COVER ALL ERRORS

The margin of error in a confidence interval covers only random sampling errors.

Practical difficulties such as undercoverage, nonresponse, or response bias are often more serious than random sampling error, yet they cannot be quantified by the margin of error. A confidence interval simply ignores such difficulties.

Recall from Chapter 6 that national opinion polls often have very low response rates (some as low as 10%) and that even small changes in the wording of questions can strongly influence results. In such cases, the margin of sampling error is probably unrealistically small. Polling organizations such as Gallup and Pew Research Center commonly include the following statement in their detailed survey results: “In addition to sampling error, question wording and practical difficulties in conducting surveys can introduce error or bias into the findings of public opinion polls.” Unfortunately, this important cautionary statement is routinely omitted by news organizations.

Experimental studies also have their set of challenges, from producing random samples to problems with attrition. Clinical trials, for instance, can enroll only volunteers and, therefore, do not truly sample from the whole population of interest. In addition, subjects may move away or decide to drop out of the study, or they may be removed because of health concerns. Any systematic bias in the attrition process would not be taken into account in the margin of error of a simple inference procedure. Always look carefully at the details of a study before you trust a reported confidence interval. Peer-reviewed scientific publications typically address these issues in detail.

APPLY YOUR KNOWLEDGE

15.4 Confidence level and margin of error. In Exercise 14.5 (page 356) you obtained a 95% confidence interval for the mean concentration of the active ingredient in a pharmaceutical product. Here are three confidence intervals from the same data with varying confidence levels: (0.83032, 0.85055), (0.83540, 0.84546), and (0.83398, 0.84689).

a. Three confidence levels were used: 80%, 90% and 99%. Which of these three intervals is the 99% confidence interval? Which one is the 80% confidence interval? Explain your reasoning.
b. Which of these three intervals is the narrowest? Explain why the narrowest interval in this set is not necessarily the most informative.

15.5 Sample size and margin of error. Consider a variable known to be Normally distributed with unknown mean μμ and known standard deviation σ=10.$\sigma = 10.$

a. What would be the margin of error of a 95% confidence interval for the population mean μμ based on a random sample of size 25? The multiplier for a zz confidence interval with a 95% confidence level is the critical value z*=1.960.$z^* = 1.960.$
b. What would be the margin of error of a 95% confidence interval for μμ based on a random sample of size 100? Based on a random sample of size 400?
c. Compare the margins of error for samples of size 25, 100, and 400. How does increasing the sample size change the margin of error of a confidence interval?

15.6 Spending on food. A 2012 Gallup survey based on a random sample of 1014 American adults indicates that American families spend, on average, $151 per week on food. The report further states that, with 95% confidence, this estimate has a margin of error of ±$7.$\pm\$7.$ Explain whether the confidence interval cited accounts for the following sources of error.

a. The poll dialed telephone numbers at random and so missed all people without phones.
b. Nonresponse—some people whose phone numbers were chosen never answered the phone in several calls or answered but refused to participate in the poll.
c. There is chance variation in the random selection of telephone numbers.
d. Respondents may not accurately assess how much they spend on food or may not answer truthfully.

HOW HYPOTHESIS TESTS BEHAVE

Hypothesis tests are widely used in reporting the results of research in many fields of applied science and in industry. New pharmaceutical products require statistically significant evidence of their effectiveness and safety before they can be marketed. Courts inquire about statistical significance in hearing class action lawsuits (many of which concern pharmaceutical products or biohazards in the workplace). Marketers want to know whether a new ad campaign significantly outperforms the old one, and medical researchers want to know whether a new therapy performs significantly better. In all these cases, statistical significance is valued because it points to an effect that is unlikely to occur simply by chance. Here are some points to keep in mind when using or interpreting hypothesis tests.

STATS IN YOUR WORLD

SHOULD TESTS BE BANNED?

Significance tests don’t tell us how large or how important an effect is. Research in psychology has emphasized tests, so much so that some think their weaknesses should ban them from use. The American Psychological Association asked a group of experts. Their response: Use anything that sheds light on your study. Use more data analysis and confidence intervals. But: “The task force does not support any action that could be interpreted as banning the use of null hypothesis significance testing or *P*-values in psychological research and publication.”

How small a PP is convincing evidence? The purpose of a hypothesis test is to evaluate, using the *P*-value, the degree of evidence provided by the sample against the null hypothesis within a specified model. But how small a *P*-value is convincing evidence against the null hypothesis? This depends to a large extent on two circumstances:

- *How plausible is* H0?H_0? If H0H_0 represents an assumption that the people you must convince have long believed to be true, strong evidence (small P) will be needed to persuade them to reject H_0.

- *What are the consequences of rejecting* H_0? If, for example, rejecting H_0 in favor of H_a means making an expensive changeover from one established therapy to a newer one, you need strong evidence that the new treatment will save or improve lives. In contrast, preliminary studies do not require taking any action beyond deciding whether to pursue a scientific line of inquiry: It makes sense to continue work in light of even weakly significant evidence.

These criteria are somewhat subjective and, therefore, open to disagreement. Historically, users of statistics have often emphasized standard levels of significance such as 5% and 1%. However, *there is no sharp border between “statistically significant” and not, only increasingly strong evidence as the P-value decreases. It makes no sense to treat* $P \leq 0.05$ *as a universal rule for what is statistically significant.* On the one hand, there is no practical distinction between the *P*-values 0.048 and 0.052. On the other hand, the *P*-value 0.004 offers clearly stronger evidence against H_0 than the *P*-value 0.04. Providing the exact *P*-value of a test, or at least its order of magnitude, allows each of us to decide individually if the evidence against H_0 is sufficiently strong.

Convincing evidence, however, is not the same as certainty. Because the decision to reject or fail to reject the null hypothesis is based on a probability value and, it could be the wrong decision. Rejecting H_0 when H_0 is actually true is called a **Type I error.** By definition, whatever significance level we choose determines the chance that a decision to reject H_0 is a Type I error. Failing to reject H_0 when H_a is actually true is called a **Type II error.** The chance that a decision to fail to reject H_0 is a Type II error depends on a number of factors, as detailed in the optional section on planning studies.

Statistical significance depends on the alternative hypothesis You may have noticed that the *P*-value for a one-sided test is typically one-half the *P*-value for the two-sided test of the same null hypothesis using the same data. The two-sided *P*-value combines two equal areas, one in each tail of a symmetric density curve. The one-sided *P*-value is just one of these areas, in the direction specified by the alternative hypothesis. It makes sense that the evidence against H0H_0 is stronger when the alternative is one-sided, because the evidence is based on the data *plus* information about the direction of possible deviations from H0.H_0. If you lack this added information, you must use a two-sided alternative hypothesis. The added information should come from existing scientific knowledge or a legitimate inquiry established *before* data collection. *Sample data is used as evidence, so it should never be the basis for determining the direction of the alternative hypothesis.*

Statistical significance depends on sample size Statistical significance says that an observed effect (the difference between the sample statistic and the assumed value of the parameter under the null hypothesis) would be unlikely to happen if the null hypothesis was true. This raises an important question: Can a tiny effect be statistically significant? Yes. To see why, think of the zz test statistic for inference about a population mean as follows:

z=x--μ0σ/n=size of the observed effectsize of chance variation

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}=\frac{\text{size of the observed effect}}{\text{size of chance variation}}$$

The numerator x--μ0$\bar{x}-\mu_0$ shows how far the data diverge from the parameter in the null hypothesis. The denominator represents the size of chance variations from sample to sample, measured by the standard deviation of x-,$\bar{x}$, and is a function of the sample size n.n. Therefore, statistical significance depends on the combination of three factors:

- The size of the observed effect (x--μ0)$(\bar{x}-\mu_0)$
- The variability of individuals in the population (σσ)

- The sample size (nn)

Because all three factors affect the *P*-value, we cannot conclude that a small *P*-value is necessarily due to a large effect size. *Even a tiny effect can be highly statistically significant if the sample size is very large, individual variations are very small, or both. Inversely, even a large effect may not reach statistical significance if the sample size is too small, the data are extremely variable, or both.* Therefore, we must always keep in mind that “failing to reject the null hypothesis” does not imply that the null hypothesis is true.

In Chapters 17 through 28, we will see several procedures to test hypotheses about various parameters under different data collection designs. The relevance of sample size and the limitations of *P*-values discussed in this chapter apply to all these procedures as well.

SAMPLE SIZE AFFECTS STATISTICAL SIGNIFICANCE

Because large random samples have small chance variation, very small deviations from the null hypothesis can be highly significant if the sample is large.

Because small random samples have a lot of chance variation, even large deviations from the null hypothesis can fail to be statistically significant if the sample is small.

EXAMPLE 15.4 Facebook’s large-scale experiment

Facebook conducted a randomized experiment to see whether emotional contagion occurs on social media, in the absence of in-person interaction. The company reduced for one week the amount of emotional (positive, negative, or neutral) content from friends in the News Feed of a random sample of 689,003 Facebook users without informing them, then monitored the users’ posts the following week.

G+
Google+
Reddit
Viber
neo
1
Instagram
Facebook
Twitter
Pinter

Twin Design/Shutterstock.com

The published findings state that “omitting emotional content reduced the amount of words the person subsequently produced, both when positivity was reduced (z=-4.78,P<0.001)$(z = -4.78, P < 0.001)$ and when negativity was reduced (z=-7.219,P<0.001).$(z = -7.219, P < 0.001)$.” The findings are highly statistically significant (P<0.001),$(P < 0.001)$, yet they correspond to a Facebook user producing “an average of one fewer emotional word, per thousand words, over the following week.”[5]

While the observed effect size is extremely small, the huge sample size helps make it highly statistically significant. What the very small *P*-values tell us here is that even something as tiny as one word in 1000, on average, would not likely be observed in a random sample of 689,003 Facebook users if manipulating the users’ News Feeds had no effect on their following posts.

Statistical significance does not imply practical importance

CAUTION

Statistical significance means “the sample showed an effect larger than would often occur just by chance if the null hypothesis was true.” It says nothing about how large the effect actually is. Likewise, it says nothing about the importance of the effect in context. *Always look at the actual size of an effect in context, as well as at its statistical significance.* A confidence interval is more helpful than a *P*-value to assess the approximate size of an effect. However, even a confidence interval means little without context. Only a solid understanding of the subject matter can tell us whether an estimated effect size is consequential and how important the finding may be.

STATISTICAL SIGNIFICANCE DOES NOT IMPLY IMPORTANCE

Statistical significance does not tell us whether an effect is large enough to be important in context. That is, **statistical significance is not the same thing as practical "significance,"** an everyday word commonly used as a synonym for "meaningful" or "important."

The randomized experiment described in Example 15.4 showed that reducing the emotional content of Facebook users' News Feed for one week resulted in an effect size on the order of one fewer emotional word per 1000 words posted over the following week. The observed effect size is clearly very small, but Facebook is used a lot by a lot of people. So what is the practical importance of this finding?

EXAMPLE 15.5 Facebook's experiment: Is the effect important?

The objective of the Facebook experiment was to study emotional contagion on social media. The response variable chosen for that purpose was the relative number of emotional words in future posts. However, posting fewer emotional words may not be the same thing as feeling emotions less intensely. Perhaps Facebook users tend to repeat words they read in other posts, simply because the words are fresh in their mind or because they seem trendy. Therefore, we cannot tell whether the observed effect truly reflects "emotional contagion" or something more akin to herd behavior, which wouldn't have much practical importance at all.

The study found a very small effect when some friends' posts containing at least one emotional word were removed. Most people (except for bullies) don't purposefully alter their posts to manipulate their friends' emotions on a large scale. Therefore, even if the effect reflects true emotional contagion, the wording of friends' posts may not impact emotions much in a realistic setting.

> However, even tiny effects can sometimes have a large impact in aggregate. Facebook has more than 1 billion active users worldwide and the company filters out all News Feeds using “a ranking algorithm that Facebook continually develops and tests in the interest of showing viewers the content they will find most relevant and engaging.” In the long run, and over a massive worldwide scale, the overall effect could be meaningful.

Example 15.5 shows that determining importance in context is a lot more complex and challenging than assessing statistical significance. Always remember that statistical significance has a very narrow definition and offers extremely limited conclusions. *A data analysis should never focus on and conclude with just a P-value or a significance level.*

What a *P*-value isn’t The *P*-value of a hypothesis test is the probability, when the population conforms to the model specified by the test assumptions, of obtaining a sample statistic at least as extreme (in the direction of the alternative hypothesis) as that actually observed *if* the null hypothesis was true. This may seem like an overly convoluted definition, but it is not. To correctly interpret a *P*-value, it is necessary to understand all aspects of this definition. Oversimplification unfortunately leads to several common and consequential mistakes.

- *The P-value is not the probability that the null hypothesis is true.* A large *P*-value says only that the sample data collected would not be surprising if the test assumptions and the null hypothesis were both correct. These data may be compatible with the population model we are investigating, but that doesn’t mean that they would be incompatible with any other model. A large *P*-value, no matter how large, is simply inconclusive.
- *The P-value does not give the probability that the data are the result of chance variations alone.* It gives the probability that the data (or data even more extreme) *could* be the result of chance variations alone *if* the test assumptions and the null hypothesis are both correct. Because sample data cannot tell us when the null

hypothesis is actually true, the *P*-value does not measure the chance that the observed effect is just random noise.

- *The P-value is not useful if you haven't checked that all the test assumptions are met.* A hypothesis test compares actual sample data in relation to a specified model for the data. This model is based on a set of assumed properties of the population and the data collection process. The null hypothesis is just one element of this model. Therefore, *a small P-value can happen either when the test assumptions are not met, when the null hypothesis is not true, or both.* This means that we can legitimately reject the null hypothesis when the *P*-value is small only if we know that the test assumptions are met.

Beware of multiple analyses The reasoning behind statistical significance works well if you decide what effect you are seeking, design a study to search for it, and use a hypothesis test to weigh the evidence you get. However, for every null hypothesis you reject because you found a small *P*-value, there is a chance that the decision is wrong, representing a Type I error. *Running one hypothesis test and reaching the 5% level of significance is reasonably good evidence that you have found something real. Running 20 tests and reaching that significance level only once is not.*

Hypothesis tests are not designed to be used blindly over and over again in hopes that something will come up statistically significant—an improper approach known as “P*P*-hacking” or “data dredging.” The concern over multiple analyses applies to confidence intervals as well. Another form of cherry picking happens when only statistically significant findings are publicly reported. The resulting publication bias undermines the entire scientific process and lowers the usefulness of a *P*-value.

If possible, replicate the study Statistical inference is a probabilistic approach that uses sample data to conclude about a population. This approach has a number of limitations, as we have discussed here. One way to improve on it is to **replicate** an entire study, from data collection to data analysis. When a study is reproduced and the findings are independently confirmed, the validity of scientific conclusions is reinforced.

APPLY YOUR KNOWLEDGE

15.7 Detecting acid rain. Emissions of sulfur dioxide by industry set off chemical changes in the atmosphere that result in “acid rain.” The acidity of liquids is measured by pH on a scale of 0 to 14. Distilled water has pH 7.0, and lower pH values indicate acidity. Typical rain is somewhat acidic, so acid rain is defined as rainfall with a pH below 5.0.

Suppose that pH measurements of rainfall on different days in a Canadian forest follow a Normal distribution with standard deviation σ=0.5.$\sigma = 0.5$. A sample of nn days finds that the mean pH is x-=4.8.$\bar{x} = 4.8$. Is this good evidence that the mean pH μμ for all rainy days is less than 5.0? The answer depends on the size of the sample. Use the *P-Value of a Test of Significance* applet or technology for your computations.

a. Enter H0:μ=5,$H_0 : \mu = 5$, Ha:μ<5,$H_a : \mu < 5$, σ=0.5,$\sigma = 0.5$, and x-=4.8.$\bar{x} = 4.8$. Then enter n=5,$n = 5$, n=15,$n = 15$, and n=40$n = 40$ one after the other, clicking “UPDATE” each time to get the three *P*-values. What are they?

b. Sketch the three Normal curves displayed by the applet, with x-=4.8$\bar{x} = 4.8$ marked on each curve. Explain why the *P*-value of the same result x-=4.8$\bar{x} = 4.8$ is smaller (more statistically significant) for larger sample sizes.

c. The corresponding 95% confidence intervals for the mean pH μμ would be 4.36 to 5.24 (for n=5$n = 5$), 4.55 to 5.05 (for n=15$n = 15$), and 4.65 to 4.95 (for n=40$n = 40$). What information is conveyed by a confidence interval that we can’t learn from the *P*-value alone?

15.8 Is it statistically significant? A study examined dairy cows during pregnancy and after giving birth (“calving”) and reported the following:

> *The effect of time was almost significant at the 5% level (P=0.058),*$(P = 0.058)$, *with the mean locomotion scores tending to be higher after calving than in late pregnancy.*[6]

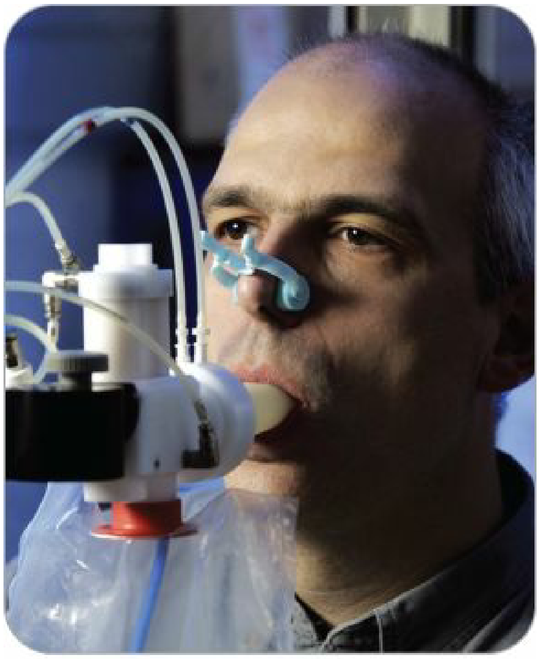
a. Why is it more informative to report the actual *P*-value than to simply reject or fail to reject H0H_0 at a chosen significance level?
b. Explain what “almost significant at the 5% level” means.

15.9 Two types of error. Your company markets a computerized medical diagnostic program used to evaluate thousands of people. The program scans the results of routine medical tests (for example, pulse rate and blood tests) and refers the case to a doctor if if finds evidence of a medical problem. The program makes a decision about each person.

a. What are the null and alternative hypotheses and the two types of error that the program can make? Describe the two types of error in terms of “false-positive” and “false-negative” test results.
b. The program can be adjusted to decrease one error probability, at the cost of an increase in the other error probability. Which error probability would you choose to make smaller, and why? (This is a matter of judgment. There is no single correct answer.)

15.10 Treating shortness of breath. Dyspnea, or shortness of breath, is a common complaint in patients with chronic obstructive pulmonary disease (COPD). It is often assessed by an FEV_1 test, which measurs the forced expiratory volume in the first second. A study examined the effect of various treatments on perceived dyspnea in patients with advanced COPD. The researchers reported that, for the 23 patients who had received 6 weeks of therapy with a long-acting bronchodilator, “there was a small, statistically insignificant, increase in FEV_1.”[7]

a. What does “statistically insignificant” mean?
b. Why is it important to know that the effect was small in size as well as statistically insignificant?



John Thys/Reporters/Science Source

15.11 Searching for ESP. A researcher looking for evidence of extrasensory perception (ESP) tests 500 subjects. Four of these subjects do significantly better (P<0.01)$(P < 0.01)$ than random guessing.

a. Is it proper to conclude that these 4 people have ESP? Explain your answer.
b. What should the researcher now do to test whether any of these 4 subjects have ESP?

DISCUSSION The scientific approach: working hypotheses, study design, data collection, statistical analysis, and discussion of findings in context

Citing a *P*-value is not the ultimate goal and focus point of a research study. In fact, a *P*-value or a confidence interval will be meaningless if they originated in a flawed design. Statistical inference is only part of the whole scientific approach. Good scientific methodology follows a multiple-step approach: Define the working hypotheses, select a suitable study design, collect and organize data, perform an appropriate statistical analysis, and discuss the findings in context along with the strengths and weaknesses of the whole process. Statistical analysis can lead to valuable conclusions only when the entire approach is sound.

A data-driven scientific inquiry starts with sound working hypotheses. These may emerge in complicated ways from the results of past studies, preliminary data, theory, or even an educated guess. The hypotheses must be specific enough to allow studies to be designed to challenge them. Simplicity is key when determining a first study design, because dealing with too many factors can hamper our ability to reach clear conclusions. While complexity is important to uncover the role of lurking variables or the existence of subgroups with different responses, it is best added one step at a time as our understanding of the topic improves. The ability to reach informative conclusions from statistical inference further depends on the quality and quantity of data, the assurance that all the conditions required for inference are met, and a comprehensive discussion of the limitations of the approach and conclusions in context.[8] Here is an outstanding example of use of the scientific method.

Uncovering the cause of peptic ulcers[9]

It was once widely assumed that stress, an anxious personality, and even spicy food caused peptic ulcers—painful lifelong lesions of the stomach or duodenum for which only temporary relief was available. Several reports had noted unusual, curved bacteria found in biopsies of patients suffering from peptic ulcers.

These were undeniable observations, but at the time, most researchers and physicians believed that nothing could survive very long in the stomach because of its extremely acidic environment (pH between 2 and 4). Thus, the reports had been generally ignored—except by Australian researchers Robin Warren and Barry Marshall. They thought that it was worth exploring a possible link between the curved bacteria and peptic ulcers, however counterintuitive it might be. They had a working hypothesis.

How can you test the hypothesis that a given bacterium causes peptic ulcers? It would obviously not be ethical to infect people with the bacteria and see if they developed a peptic ulcer. (Interestingly, that's exactly what Marshall did to himself later on, after he and Warren had developed a treatment against the bacteria.) The only real research option at this stage was to conduct an observational study. Warren and Marshall examined biopsies of the stomach lining from a random sample of patients given an endoscopy referral for any number of problems along the upper digestive tract. Statistical analysis showed that the bacteria were found significantly more often in patients with peptic ulcers than in patients with other digestive disorders. This clearly established a link between the curved bacteria and peptic ulcers, an important preliminary step in determining a possible causal effect.

Dr. Gary Gaugler/Science Source

Establishing causality would need direct experimentation, however, and ethically this would require offering possible treatments. If the bacteria caused the peptic ulcers, then a treatment killing them or inhibiting their growth should relieve the symptoms. This logic gave Warren and Marshall a second hypothesis to test. They designed a double-blind experiment in which patients with an established diagnosis of peptic ulcer received one of several treatments or a placebo. They found that patients whose treatment completely eradicated the bacteria did significantly better over both the short term and the long term, with very few relapses even many years later. The experiment showed that the presence of the bacteria was indeed responsible, at least in part, for the peptic ulcers. It also showed that the majority of peptic ulcers could be successfully treated, thereby preventing a lifetime of suffering for these patients.

How these curved bacteria, *Helicobacter pylori*, could survive in the acidic environment of the stomach and duodenum was still a mystery that needed to be resolved to undeniably establish the bacterial role in peptic ulcers. A study showed that *H. pylori* produces urease, an enzyme that had already been found in the gastric mucosa of animals. Urease breaks down urea into carbon dioxide and ammonia, resulting in a higher (more basic) pH in solution. Marshall and colleagues hypothesized that *H. pylori* might use urease to survive in the extremely acidic gastric environment by neutralizing the pH locally. To test this hypothesis, they compared the survival, in environments with various pH levels, of urease-producing *H. pylori* and two other types of bacteria with no or limited capacity to produce urease. In the absence of urea, none of the three bacteria types survived in significant numbers at pH below 3.

When urea was added to the solution, however, *H. pylori* survived in significant numbers at pH levels as low as 1.5, whereas the other two types still couldn’t survive below pH 3.

This experiment showed that *H. pylori* possessed a mechanism enabling it to survive in the stomach long enough to have a lasting impact. It also paved the way for a new, fast, and noninvasive method to assess the presence of *H. pylori* in patients with gastric problems simply by having patients swallow a solution of carbon-labeled urea and monitoring signs of labeled carbon dioxide in their breath shortly afterward.

These are only three of the many studies Warren and Marshall conducted on the link between *H. pylori* and peptic ulcers over two decades. In 2005, they were awarded the Nobel Prize in Medicine “for their discovery of the bacterium *Helicobacter pylori* and its role in gastritis and peptic ulcer disease.” More about this topic can be found on the Nobel Prize website at nobelprize.org/nobel_prizes/medicine/laureates/2005.

APPLY YOUR KNOWLEDGE

15.12 Mammary artery ligation. Angina is the severe pain caused by inadequate blood supply to the heart. Perhaps we can relieve angina by tying off the mammary arteries to force the body to develop other routes to supply blood to the heart. When surgeons tried this procedure, called “mammary artery ligation,” patients reported a statistically significant reduction in angina pain. A randomized comparative experiment later showed that ligation was no more effective than a sham operation (placebo). Surgeons abandoned the procedure at once.[10] Which aspect of the scientific approach presented in the preceding Discussion was problematic, resulting in a misleading interpretation of the statistically significant reduction in angina pain?

15.13 Convincing evidence. We discussed earlier how small a *P*-value should be so we can reject the null hypothesis. One factor mentioned was how plausible the null hypothesis is, such that rejecting a null hypothesis that is generally considered true may demand especially strong evidence (small significance level). Which of the studies described in the preceding Discussion would demand the smallest significance level for this particular reason? Explain your reasoning.

PLANNING STUDIES: SELECTING AN APPROPRIATE SAMPLE SIZE[i]

A wise user of statistics never plans a sample or an experiment without also planning the inference. The number of observations is a critical part of planning a study. Larger samples give smaller margins of error in confidence intervals and ensure that hypothesis tests are better able to detect effects in the population. But taking observations costs both time and money. How many observations are enough? We will look at this question in the context of one-sample zz procedures for a population mean because the computations are easiest in this setting. Planning for a study is just as relevant for other inference procedures, although in their case specialized software is often needed.

Let’s start with confidence intervals. You can arrange to have both high confidence and a small margin of error by taking enough observations. The margin of error of the confidence interval for the mean of a Normally distributed population is m=z*σ/n.$m = z^*\sigma/\sqrt{n}$. To obtain a desired margin of error m,m, find the value of z*z^* for your desired confidence level, and solve for the sample size n.n. Here is the result.

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The confidence interval for the mean of a Normal population will have a specified margin of error mm when the sample size is

n=z*σm2

$$n = \left(\frac{z^*\sigma}{m}\right)^2$$

Notice that it is the size of the sample that determines the margin of error. The size of the population does not influence the sample size we need. (This is true as long as the population is much larger than the sample.)

EXAMPLE 15.6 How many observations?

Example 14.10 (page 368) reports a study of the mean body temperature of healthy adults. We know that the population standard deviation is σ=0.6$\sigma = 0.6$ degree Fahrenheit (°F).$(^\circ\text{F})$. We want to estimate the mean body temperature μμ for healthy adults within ±0.05 °F$\pm 0.05\ ^\circ\text{F}$ with 95% confidence. How many healthy adults must we measure?

The desired margin of error is m=0.05 °F.$m = 0.05\ ^\circ\text{F}$. For 95% confidence, the critical value is z*=1.96$z^* = 1.96$ (from software or Table C). We know that σ=0.6 °F.$\sigma = 0.6\ ^\circ\text{F}$. Therefore,

n=z*σm2=1.96×0.60.052=553.2

$$n = \left(\frac{z^*\sigma}{m}\right)^2 = \left(\frac{1.96 \times 0.6}{0.05}\right)^2 = 553.2$$

Because a sample of 553 healthy adults will give a slightly larger margin of error than desired, and a sample of 554 healthy adults a slightly smaller margin of error, we must recruit 554 healthy adults for the study. *Always round up to the next higher whole number when finding a large enough n.*

Now for hypothesis testing. How large a sample should we take when we plan to carry out a significance test? We know that if our sample is too small, even large effects in the population will often fail to give statistically significant results. Here are the questions we must answer to decide how large a sample we must take:

Significance level. How much protection do we want against getting a statistically significant result from our sample when there really is no effect in the population?

Effect size. How large an effect in the population is important in practice?

Power. How confident do we want to be that our study will detect an effect of the size we think is important?

The three boldface terms are statistical shorthand for three pieces of information. *Power* is a new idea.

EXAMPLE 15.7 Sweetening colas: planning a study

A quality control engineer plans a study to assess the impact of storage on the sweetness of a new cola. Ten trained tasters are available to rate the cola's sweetness on a 10-point scale before and after storage, so that each taster's judgment of loss of sweetness can be assessed. Industry records indicate that sweetness loss scores vary from taster to taster according to a Normal distribution with standard deviation $\sigma=1$.$\sigma = 1$. To see if the taste test suggests that the cola does lose sweetness, the engineer will test

$$H_0: \mu = 0 \quad H_a: \mu > 0$$

$$H_0 : \mu = 0$$
$$H_a : \mu > 0$$

Are 10 tasters enough, or should more be used?

Dadang Tri/Bloomberg via Getty Images

Significance level. Requiring significance at the 5% level is enough protection against declaring there is a loss in sweetness when no change would actually be found if we could look at the entire population. It means that when there is no change in sweetness in the population, 1 out of 20 samples of tasters will wrongly find a statistically significant loss.

Effect size. A mean sweetness loss of 0.8 point on the 10-point scale will be noticed by consumers, so it is important in practice.

Power. The engineer needs to be 90% confident that the test will detect a mean loss of 0.8 point in the population of all tasters. A significance level of 5% is standard in the food industry for detecting an effect. Thus, the objective is to have probability at least 0.9 that a test at the $\alpha = 0.05$ level will reject the null hypothesis $H_0 : \mu = 0$ when the true population mean is $\mu = 0.8$.

The probability that the test successfully detects a sweetness loss of the specified size is the *power* of the test. You can think of tests with high power as being highly sensitive to deviations from the null hypothesis. In Example 15.7, the engineer decided on power 90% when the truth about the population is $\mu = 0.8$.

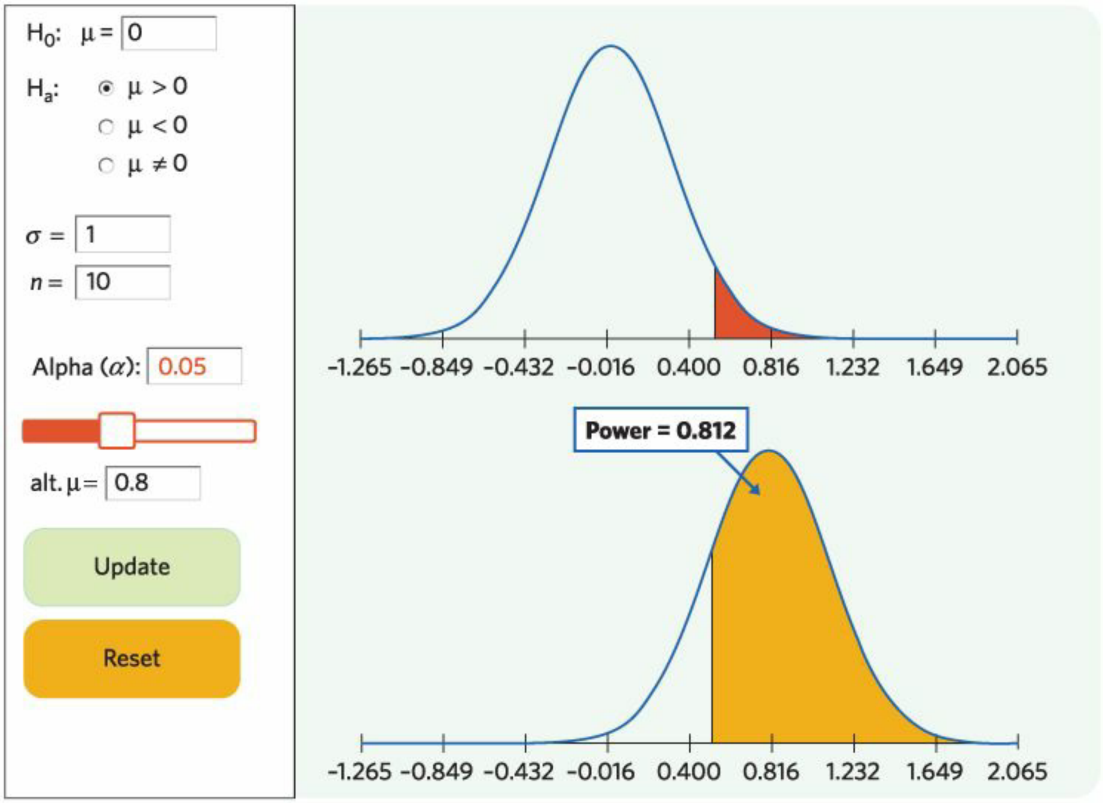
POWER

The **power** of a test against a specific alternative is the probability that the test will reject H_0 at a chosen significance level α when the specified alternative value of the parameter is true.

For most statistical tests, calculating power is a job for specialized statistical software. Finding it for the z test is easier, but we will nonetheless skip the details (using an applet and software instead). Exercise 15.42 walks you through the details of the calculations.

EXAMPLE 15.8 Sweetening colas: finding power

Finding the power of the z test is less challenging than most other power calculations because it requires only a Normal distribution probability calculation. The *Statistical Power* applet does this and illustrates the calculation with Normal curves. Enter the information from Example 15.7 into the applet: hypotheses, population standard deviation $\sigma = 1$, sample size $n = 10$, significance level $\alpha = 0.05$, and alternative $\mu = 0.8$. Click “UPDATE.” The applet output appears in Figure 15.2.



H_0: $\mu =$ 0

H_a: ● $\mu > 0$ ○ $\mu < 0$ ○ $\mu \neq 0$

$\sigma =$ 1

$n =$ 10

Alpha (α): 0.05

alt. $\mu =$ 0.8

Update
Reset
-1.265 -0.849 -0.432 -0.016 0.400 0.816 1.232 1.649 2.065
Power = 0.812
-1.265 -0.849 -0.432 -0.016 0.400 0.816 1.232 1.649 2.065

Figure 15.2

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 15.2 Output from the Statistical Power applet for the cola study with σ=1, n=10,

$$\sigma = 1, n = 10,$$

and significance level 0.05. The power of the test H0:μ=0

$$H_0 : \mu = 0$$

against the specific alternative Ha:μ=0.8

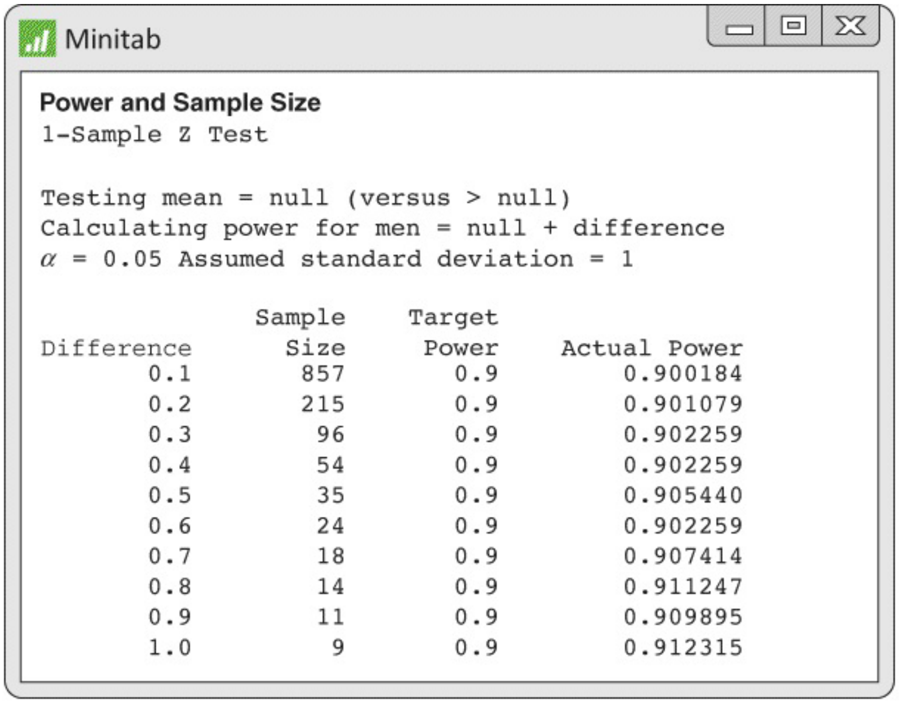
$$H_a : \mu = 0.8$$

is 0.812.

The two Normal curves in Figure 15.2 show the sampling distribution of x-$\bar{x}$ under the null hypothesis μ=0$\mu = 0$ (top) and under the specific alternative μ=0.8$\mu = 0.8$ (bottom). The curves have the same shape because σ/n$\sigma/\sqrt{n}$ does not change. The shaded region at the right of the top curve has area 0.05. It marks off values of x-$\bar{x}$ that are statistically significant at the α=0.05$\alpha = 0.05$ level. The lower curve shows the probability of these same values when μ=0.8.$\mu = 0.8.$ This area is the power, 0.812.

The power of the test against the specific alternative μ=0.8$\mu = 0.8$ is 0.812. That is, the test will reject H0H_0 about 81% of the time when this alternative is true. Therefore, 10 observations are too few to give power 90%.

Some software programs automate the process of selecting the desired sample size for a given power. Figure 15.3 gives Minitab’s output for the one-sided zz test aiming to have power 0.9 against several specific alternatives at the 5% significance level when the population standard deviation is σ=1.$\sigma = 1.$ In this output, “Difference” is the difference between the null hypothesis value μ=0$\mu = 0$ and the alternative we want to detect. In other words, it is the effect size. The “Sample Size” column shows the smallest number of observations needed for power 0.9 against each effect size.



```
Power and Sample Size
1-Sample Z Test

Testing mean = null (versus > null)
Calculating power for men = null + difference
α = 0.05 Assumed standard deviation = 1

              Sample    Target
Difference      Size     Power    Actual Power
       0.1       857       0.9        0.900184
       0.2       215       0.9        0.901079
       0.3        96       0.9        0.902259
       0.4        54       0.9        0.902259
       0.5        35       0.9        0.905440
       0.6        24       0.9        0.902259
       0.7        18       0.9        0.907414
       0.8        14       0.9        0.911247
       0.9        11       0.9        0.909895
       1.0         9       0.9        0.912315
```

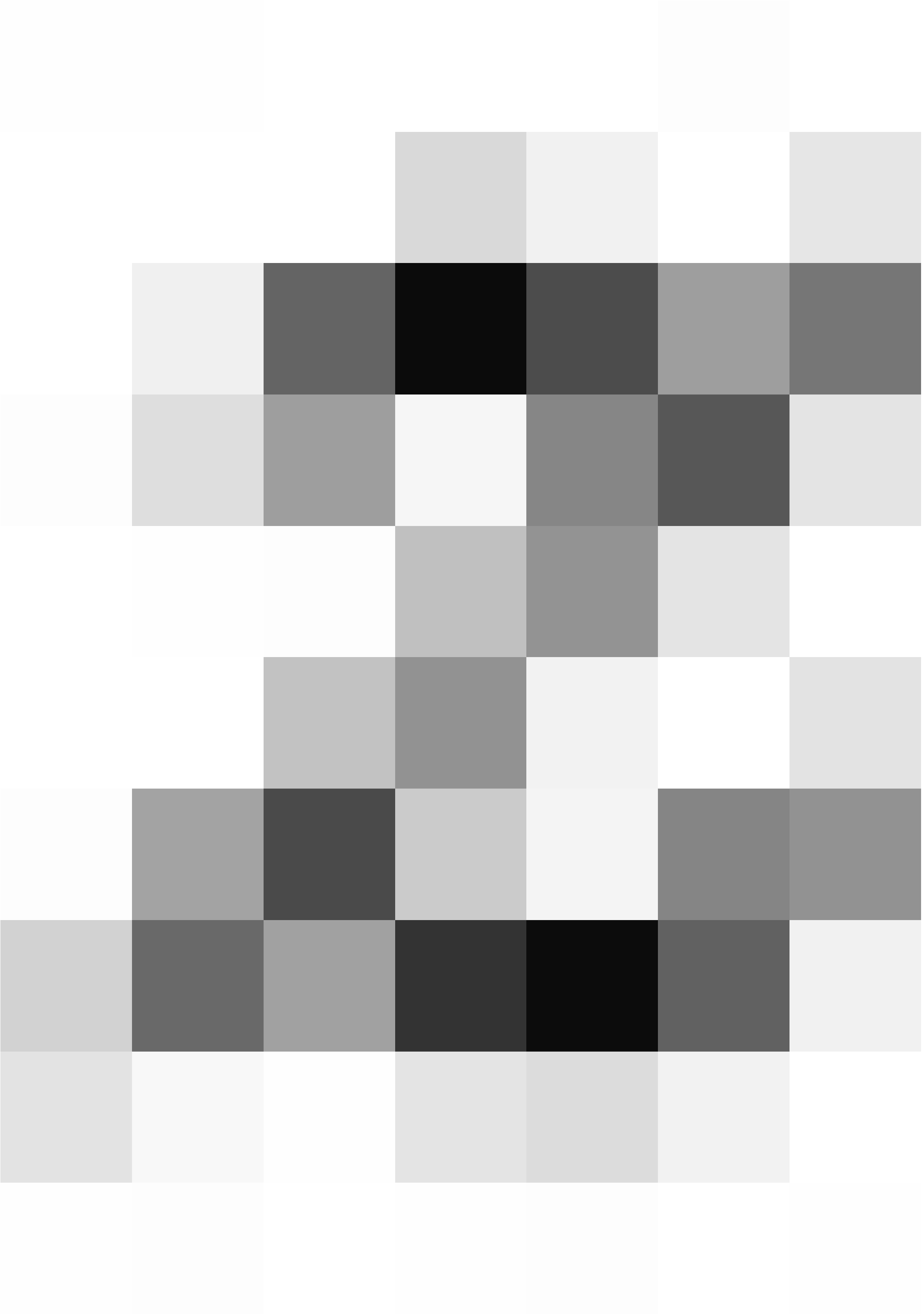
Figure 15.3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

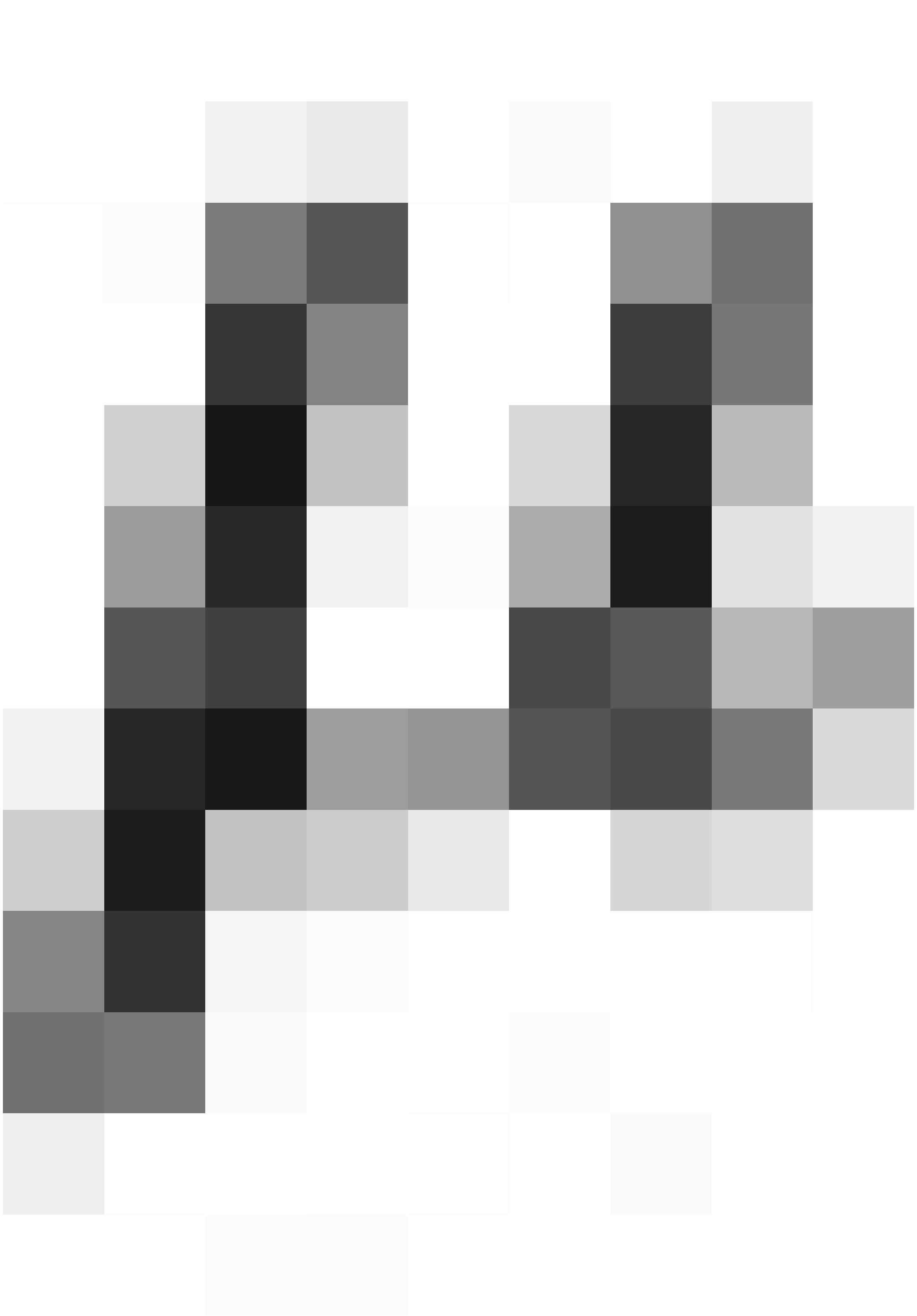
FIGURE 15.3 Output from Minitab for the cola study with σ=1

$$\sigma=1$$

and significance level 0.05. The sample size required to reach a target power of 0.9 for the one-sided z



test aiming for various effect sizes, here labeled “difference” (the difference between μ



under H0

> H_0
>
> and Ha).
>
> H_a).

The software output in Example 15.8 shows clearly that, if the study aims to detect smaller effects, then larger samples are required to reach the same 90% power. Here is an overview of influences on the sample size needed:

- If you insist on a smaller significance level (such as 1% rather than 5%), you will need a larger sample. A smaller significance level requires stronger evidence to reject the null hypothesis.
- If you insist on higher power (such as 90% rather than 80%), you will need a larger sample. Higher power gives a better chance of detecting an effect when it is really there.
- At any significance level and desired power, a two-sided alternative requires a larger sample than a one-sided alternative.
- At any significance level and desired power, aiming to detect a small effect requires a larger sample than aiming to detect a large effect.

Planning a serious statistical study always requires an answer to the question “How large a sample do I need?” If you intend to test the hypothesis $H_0 : \mu = \mu_0$ about the mean μ of a population, you need at least a rough idea of the value of the population standard deviation σ and the magnitude of the deviation $\mu - \mu_0$ of the population mean from its hypothesized value that you want to be able to detect. More elaborate settings, such as comparing the mean effects of several treatments, require more elaborate advance information. You can leave the details to experts, but you should understand the idea of power and the factors that influence how large a sample you need.

Calculating the power of a test requires a fixed level of significance, such as $\alpha = 0.05$. In practice, however, it is better to think in terms of *P*-values rather than a fixed level α. To effectively plan a new study you should find the power for a range of sample sizes and effect sizes to get a full picture of how the test will behave.

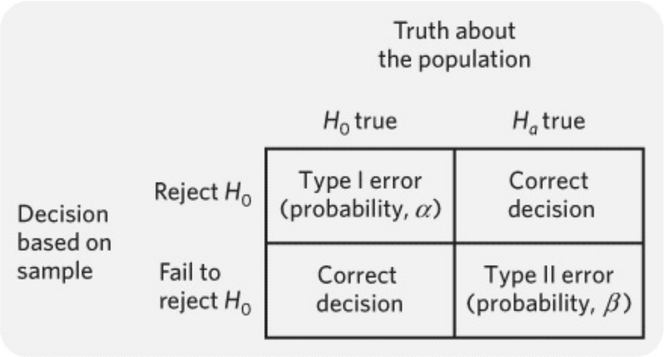
Power, Type I error, and Type II error in hypothesis tests. We can assess the performance of a test by giving two probabilities: the significance level α and the power for an alternative that we want to be able to detect. The significance level of a test is the probability of making the *wrong* decision when the null hypothesis is true. The power for a specific alternative is the probability of making the *right* decision when that alternative is true. We can just as well describe the test by giving the probability of a *wrong* decision under both conditions.

PROBABILITY OF TYPE I AND TYPE II ERRORS

The **significance level** αα of any fixed level test is the probability of a Type I error.

The **power** of a test against *any specific* alternative is 1 minus the probability of a Type II error for *that* alternative: power=1-β.$\text{power} = 1 - \beta.$

The possibilities are summed up in Figure 15.4. When H0H_0 is true, our decision is correct if we fail to reject H0,$H_0,$ but it is a Type I error if we reject H0.$H_0.$ There is only one possible value for the parameter of interest when H0H_0 is true, so the probability of a Type I error is just the significance level α,$\alpha,$ regardless of any other factor.



		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error (probability, α)	Correct decision
	Fail to reject H_0	Correct decision	Type II error (probability, β)

Figure 15.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 15.4 The two types of error in testing hypotheses and their probability.

When HaH_a is true, our decision is correct if we reject H0,$H_0,$ but it is a Type II error if we fail to reject H0.$H_0.$ However, if H0H_0 is not true, then we do not have an exact value for the parameter of interest; it could be anything consistent with Ha.$H_a.$ This is why the probability of a Type II error is different for different “effect sizes.”

EXAMPLE 15.9 Sweetening colas: finding error probabilities

Because the probabilities of the two types of error are just a rewording of significance level and power, we can see from Figure 15.2 what the error probabilities are for the test in Example 15.8.

P(Type I error)=P(reject H0when in fact μ=0)=significance level α=0.05P(Type II error)=P(fail to reject H0when in fact μ=0.8)=1-power=1-0.812=0.188

$$\begin{aligned} P\,(\text{Type I error}) &= P\,(\text{reject } H_0 \text{when in fact } \mu = 0) \\ &= \text{significance level } \alpha = 0.05 \end{aligned}$$

$$\begin{aligned} P\,(\text{Type II error}) &= P\,(\text{fail to reject } H_0 \text{when in fact } \mu = 0.8) \\ &= 1 - \text{power} = 1 - 0.812 = 0.188 \end{aligned}$$

The two Normal curves in Figure 15.2 are used to find the probabilities of a Type I error (top curve, μ=0$\mu = 0$) and of a Type II error (bottom curve, μ=0.8$\mu = 0.8$).

APPLY YOUR KNOWLEDGE

15.14 Estimating mean IQ. IQ scores are typically Normally distributed within a homogeneous population. How large a sample of schoolgirls in Example 14.3 (page 354) would be needed to estimate the mean IQ score μμ within ±5± 5 points with 99% confidence, assuming that σ=15$\sigma = 15$ in this population? (Note that the zz multiplier for a 99% confidence level is z*=2.576.$z^* = 2.576.$)

15.15 Pharmaceutical production. Exercise 14.5 (page 356) described the manufacturing process of a pharmaceutical product. Repeated measurements follow a Normal distribution with mean μμ equal to the true product concentration, and with standard deviation σ=0.0068 g/l. $\sigma = 0.0068\,\text{g/l}.$

a. How many measurements would be needed to estimate the true concentration within ±0.001 g/l$\pm 0.001\,\text{g/l}$ with 95% confidence?

b. How many measurements would be needed to estimate the true concentration within ±0.001 g/l$\pm 0.001\,\text{g/l}$ with 99% confidence? What is the implication of choosing a higher confidence level?

15.16 Thinking about power. An environmentalist group is interested in the following hypotheses about the mean oxygen content (in milligrams per liter, mg/l) of a stream:

$$H0: μ=5 \text{ mg/l} \quad \text{versus} \quad Ha: μ<5 \text{ mg/l}$$

$H_0 : \mu = 5 \text{ mg/l} \text{ versus } H_a : \mu < 5 \text{ mg /l}$

A liter of water will be collected from each of 45 random locations along the stream. Suppose that dissolved oxygen varies among sampling locations according to a Normal distribution with σ=0.92 mg/l.$\sigma = 0.92 \text{ mg /l.}$ A statistician says that the power of the zz test with α=0.05$\alpha = 0.05$ against the alternative that the true mean oxygen content is μ=4.75 mg/l$\mu = 4.75 \text{ mg /l}$ is 0.57.

a. Explain in simple language what “power=0.57”“$\text{power} = 0.57$” means.
b. If we decide to collect water from more than 45 locations but change nothing else in this study, does the power increase or decrease?
c. If we decide to use α=0.10$\alpha = 0.10$ instead of α=0.05$\alpha = 0.05$ but change nothing else, does the power increase or decrease?
d. If we shift our interest to the alternative μ=4.5 mg/l$\mu = 4.5 \text{ mg /l}$ but change nothing else, does the power increase or decrease?

15.17 Power and error probabilities. Consider two distinct situations.

a. A statistical test at significance level α=0.05$\alpha = 0.05$ has power 0.78. What are the probabilities of Type I and Type II errors for this test?
b. A statistical test at the α=0.01$\alpha = 0.01$ level has probability 0.14 of making a Type II error when a specific alternative is true. What is the power of the test against this alternative?

STATS IN YOUR WORLD

FISH, FISHERMEN, AND POWER

Are the stocks of cod in the ocean off eastern Canada declining? Studies over many years failed to find significant evidence of a decline. These studies had low power—that is, they might fail to find a decline even if one was present.

When it became clear that the cod were vanishing, quotas on fishing ravaged the economy in parts of Canada. If the earlier studies had had high power, they would likely have seen the decline. Quick action might have reduced the economic and environmental costs.

CHAPTER 15 SUMMARY

- A specific confidence interval or test is correct only under specific conditions. The most important conditions concern the method used to produce the data. Other factors such as the shape of the population distribution may also be important.
- Whenever you use statistical inference, you are acting as if your data are a probability sample or come from a randomized comparative experiment.
- Always perform data analysis before inference to detect outliers or other problems that would make inference untrustworthy.
- Other things being equal, the **margin of error** of a confidence interval gets smaller as
  - The confidence level *C* decreases,
  - The population standard deviation σσ decreases, and
  - The sample size nn increases.
- The margin of error in a confidence interval accounts for only the chance variation due to random sampling. In practice, errors due to nonresponse or undercoverage are often more serious.
- There is no universal rule for how small a *P*-value is convincing. Beware of placing too much weight on traditional significance levels, such as α=0.05.$\alpha = 0.05$.
- Very small effects can be highly significant (small PP) when a test is based on a large sample. A statistically significant effect need not be practically important. Use confidence intervals to estimate the actual values of parameters and interpret them in context.
- Lack of statistical significance does not imply that H0H_0 is true. Even a large effect can fail to be statistically significant when a test is based on a small sample, the data are very variable, or both.
- Smart experimenters plan their studies with inference in mind. In

particular, it is good practice to determine what sample size will allow successful inference.

- The sample size required to obtain a confidence interval with specified margin of error mm for a Normal mean can be computed before collecting data as

  n=z*σm2

  $$n = \left(\frac{z^*\sigma}{m}\right)^2$$

  where z*z^* is the critical value for the desired level of confidence. Always round nn up when you use this formula.
- The **power** of a hypothesis test measures its ability to detect when a specific alternative hypothesis is true. The power of a test against a specific alternative is the probability that the test will reject H0H_0 when that alternative is true.
- Increasing the size of the sample increases the power of a hypothesis test. Using statistical applets or software, you can find the sample size required to achieve a desired test power before actually collecting data.
- We can describe the performance of a test at fixed significance level αα by giving the probabilities of two types of error. A **Type I error** occurs if we reject H0H_0 when it is actually true. A **Type II error** occurs if we fail to reject H0H_0 when actually HaH_a is true.
- In a fixed level αα hypothesis test, the significance level αα is the probability of a Type I error, and the power against a specific alternative is 1 minus the probability of a Type II error for that alternative.
- When many hypothesis tests are run on the same data source, some will probably be statistically significant by chance alone, even if all the null hypotheses are all true; that is, they would be Type I errors.

THIS CHAPTER IN CONTEXT

This chapter expands on Chapter 14 to address the implications and limitations of statistical inference in the context of inference for a population mean under simple conditions. As with Chapter 14, what you have learned here is also broadly relevant to all the chapters on statistical inference in the remainder of this book, even if the specific computations and applications vary with each inference procedure.

The limitations discussed here with respect to interpreting inference results draw heavily on everything you learned in Chapters 6 and 7 about the challenges of data collection. Concerns about how the data were collected should lead you to question the applicability or even the validity of the conclusions from any statistical inference based on these data.

The techniques of descriptive statistics you learned in Chapters 1 and 2 are also helpful in determining whether some data are adequate for the purpose of statistical inference. You will find that these techniques are even more critically important once we step away from the simple conditions considered in Chapters 14 and 15.

CHECK YOUR SKILLS

15.18 The most important condition for drawing sound conclusions from statistical inference is usually

a. that the data can be thought of as a random sample from the population of interest.

b. that the population distribution is exactly Normal.

c. that no calculation errors are made in the confidence interval or test statistic.

15.19 The coach of a college basketball team records the resting pulse rates of the team members. A confidence interval for the mean resting pulse rate of all college-age adults based on these data is of little use because

a. the number of team members is small, so the margin of error will be large.

b. many of the students in the team will probably refuse to respond.

c. the college students in the basketball team can’t be considered a random sample from the population.

15.20 You turn your browser to a website featuring a Quick Vote poll, which allows site visitors to choose an answer to the question of the day. You view yesterday’s poll results, which are based on 26,494 responses. You should refuse to calculate any 95% confidence interval based on this sample because

a. yesterday’s responses are meaningless today.

b. inference from a voluntary response sample can’t be trusted.

c. the sample is too large.

15.21 Many sample surveys use well-designed random samples, but half or more of the original sample can’t be contacted or refuse to take part. Any errors due to this nonresponse

a. have no effect on the accuracy of confidence intervals.

b. are included in the announced margin of error.

c. are in addition to the random variation accounted for by the announced margin of error.

15.22 Which of the following questions does a hypothesis test evaluate?

a. The adequacy of the study design

b. Chance variations in random sampling

c. The importance of an observed effect

15.23 Here’s a quotation from a medical journal: “An uncontrolled experiment in 17 women found a significantly improved mean clinical symptom score after treatment. Methodologic flaws make it difficult to interpret the results of this study.” The authors of this paper are skeptical about the significant improvement because

a. there is no control group, so the improvement might be due to the placebo effect or to the fact that many medical conditions improve over time.

b. the *P*-value given was $P=0.03$,$P = 0.03$, which is too large to be convincing.

c. the response variable might not have an exactly Normal distribution in the population.

15.24 Vigorous exercise helps people live several years longer, on average. Whether mild activities like slow walking extend life, however, is not clear. Suppose that the added life expectancy from regular slow walking is just 2 months. A statistical test is more likely to find a significant increase in mean life expectancy if

a. it is based on a very large random sample.

b. it is based on a very small random sample.

c. The size of the sample doesn’t have any effect on the statistical significance of the test.

15.25 A medical experiment compared the herb echinacea with a placebo for preventing colds. One response variable was “volume of nasal secretions” (if you have a cold, you blow your nose a lot). We consider the average volume of nasal secretions in people without colds to be $\mu=1$.$\mu = 1$.

An increase to μ=3 $\mu = 3$ indicates a cold. In testing H0:μ=1 $H_0 : \mu = 1$ versus Ha:μ>1, $H_a : \mu > 1,$ a Type I error would be

a. rejecting H0 H_0 when μ=1 $\mu = 1$ is true.
b. rejecting H0 H_0 when μ=3 $\mu = 3$ is true.
c. failing to reject H0 H_0 when μ>1 $\mu > 1$ is true.

15.26 (Optional) In Exercise 15.25, the power of the test against the specific alternative μ=3 $\mu = 3$ is

a. the probability that the test rejects H0 H_0 when μ=1 $\mu = 1$ is true.
b. the probability that the test rejects H0 H_0 when μ=3 $\mu = 3$ is true.
c. the probability that the test fails to reject H0 H_0 when μ=3 $\mu = 3$ is true.

15.27 (Optional) A laboratory scale is known to have a standard deviation of σ=0.001 $\sigma = 0.001$ gram in repeated weighings. Scale readings in repeated weighings are Normally distributed, with mean equal to the true weight of the specimen. How many times must you weigh a specimen on this scale to get a margin of error no larger than ±0.0005 ± 0.0005 with 95% confidence?

a. 4 times
b. 15 times
c. 16 times

CHAPTER 15 EXERCISES

15.28 Asking about smoking habits. A 2012 Gallup survey of 1014 American adults asked participants about their smoking habits. Gallup has been conducting this survey yearly since 1944. The 2012 report was titled "In U.S., Smokers Light Up Less Than Ever," reflecting the fact that only 1% of smokers in this latest survey stated that they smoke more than a pack a day. Gallup added the following comment to its report:

> *It is possible that the decline in reports of smoking is the result of respondents' awareness that smoking is socially undesirable. Therefore, respondents may aim to present themselves in the best possible light to the interviewer and underestimate the amount they truly smoke.*

What type of bias is Gallup describing? Would it be included in the margin of error of a 95% confidence interval for the proportion of all adult smokers who smoke more than a pack a day?

15.29 Sensitive questions. The National AIDS Behavioral Surveys asked a random sample of 2673 adult heterosexuals how many sexual partners they had in the past year. Why is an estimate based on these findings likely to be biased? Does the margin of error of a 95% confidence interval for the mean number of sexual partners allow for this bias?

15.30 Saccharin no longer listed as a possible human carcinogen. In its 11th edition, the U.S. *Report on Carcinogens* deleted saccharin from the list of possible human carcinogens. Earlier carcinogenicity experiments had found rare instances of bladder tumors among male rats under very specific circumstances. Female rats never exhibited this sensitivity to saccharin, and neither did male or female mice. The toxicity was later explained by the unique metabolic pathway of certain male rats only. Years of observational studies in humans failed to show any sign of carcinogenicity from the regular consumption of saccharin.[11]

a. What was the assumption made when saccharin was first placed on the list of possible human carcinogens based on the rat data? Why do you think this assumption was made?
b. Which simple condition is not met when applying animal research to humans? Explain why confirmation of the unique physiology of certain male rats led to the removal of saccharin from the list of possible human carcinogens.

15.31 When to use pacemakers. A medical panel prepared guidelines for when cardiac pacemakers should be implanted in patients with heart problems. The panel reviewed a large number of medical studies to judge the strength of the evidence supporting each recommendation. For each recommendation, they ranked the evidence as level A (strongest), B, or C (weakest). Here, in scrambled order, are the panel's descriptions of the three levels of evidence.[12] Which is A, which B, and which C? Explain your ranking.

> *Evidence was ranked as level _______ when data were derived from a limited number of trials involving comparatively small numbers of patients or from well-designed data analysis of nonrandomized studies or observational data registries.*
>
> *Evidence was ranked as level _______ if the data were derived from multiple randomized clinical trials involving a large number of individuals.*
>
> *Evidence was ranked as level _______ when consensus of expert opinion was the primary source of recommendation.*

15.32 What distinguishes schizophrenics? A group of psychologists once measured 77 variables on a sample of people with schizophrenia and a sample of people without schizophrenia. They compared the two samples using 77 separate hypothesis tests. Two of these tests were statistically significant at the 5% level. Suppose that there is actually no difference in any of the variables between people with and without schizophrenia, so that all 77 null hypotheses are true.

a. What is the probability that 1 specific test shows a difference that is statistically significant at the 5% level?

b. Why is it not surprising that 2 of the 77 tests were statistically significant at the 5% level?

15.33 Adding fluoride to tap water. The widespread addition of fluoride to tap water in the United States and other countries has been credited with a drastic decrease in tooth decay. However, there is always some concern over possible harmful effects of any kind of treatment. Most studies so far have failed to show a negative impact of water fluoridation on health, but one study did report a significant decrease in IQ among individuals exposed to extra fluoride as children compared with individuals not exposed to this treatment. Commenting on this study in a news article on National Public Radio, Dr. Myron Allukian of the Harvard School of Dental Medicine said that the effect on IQ was small, accounting for only about a half-point difference. He further explained that a half-point difference in IQ is meaningless, just like a difference in adult heights of half a millimeter.[13]

a. Why is it important to know the size of an effect and not just that the effect was statistically significant?
b. How can an effect be both small and statistically significant?

15.34 Treating insomnia. Belsomra is a drug prescribed for the treatment of insomnia. A randomized, placebo-controlled clinical trial found that patients assigned to take Belsomra for three months fell asleep 35 minutes faster, on average, than they did before the study started—a statistically significant improvement. By comparison, patients assigned to take a placebo for three months fell asleep 27 minutes faster than they did originally. The difference between the two groups was statistically significant (reported as $P<0.01$).$P < 0.01$).[14]

a. Would you consider an effect size of falling asleep 35 minutes sooner important in the context of patients suffering from insomnia? Explain your answer.
b. The difference between taking Belsomra and taking a placebo was, on average, falling asleep 8 minutes faster with Belsomra. Would you consider an effect size of 8 minutes important in the context of patients suffering from insomnia? Explain your answer.

c. Which limitation of statistical significance does this study illustrate?

15.35 Why are larger samples better? Statisticians prefer reasonably large samples. Describe briefly the effect of increasing the size of a sample (or the number of subjects in an experiment), if all facts about the population remain unchanged, on each of the following:

a. The margin of error of a 95% confidence interval
b. The *P*-value of a test, when $H_0$$H_0$ is false
c. The impact of an outlier in the sample data on a confidence interval or a *P*-value

15.36 Cell phones and brain cancer. The INTERPHONE study is an international case-control study of individuals with brain cancer (2708 with glioma and 2409 with meningioma) and similar controls without brain cancer. The researchers compared the rates of each of these 2 types of brain cancer among subgroups of individuals, based on a number of factors: amount and history of regular cell phone use (8 levels), cell phone type (2 levels), and brain location of the cancer (5 levels). In all, the same inference procedure was therefore performed 160 times. The statistical analysis showed a small, significant "protective" effect of cell phone use in 7 groups, and a moderate, significant "harmful" effect of cell phone use in 2 groups (the researchers also noted that some individuals in these two groups provided unrealistically high reports of cell phone use).[15] Explain why, using a significance level of 0.05, it shouldn't be surprising that some of the 160 identical inference procedures were statistical significant. How does this affect your interpretation of the study findings?

The following exercises concern the optional material on planning studies for successful inference, choosing a sample size, and power calculations.

15.37 Aspirin tablets: how large a sample? Suppose we know that the aspirin content of aspirin tablets in Example 14.6 (page 357) follows a Normal distribution with standard deviation $\sigma = 5$ mg.$\sigma = 5$ mg.

a. You must verify the aspirin content of tablets produced in a day to within ±1±1 mg with 99% confidence. How large a sample of aspirin tablets from the daily production do you need?

b. If you needed to obtain only a 95% confidence interval with a margin of error no larger than 1 mg, how large a sample of aspirin tablets from the daily production would you need instead?

15.38 Deer mice. Deer mice (*Peromyscus maniculatus*) are small rodents native to North America. Their body lengths (excluding tail) are known to vary approximately Normally, with mean $\mu=86$$\mu = 86$ mm and standard deviation $\sigma=8$$\sigma = 8$ mm.[16] You decide to study the effect of vitamin supplements on deer mouse body length. First, however, you want to figure out how many deer mice you would need for successful inference about the mean $\mu$$\mu$ in the population of all deer mice that would ever receive vitamin supplements. How many deer mice would be needed to estimate $\mu$$\mu$ to within $\pm 2$$\pm 2$ mm with 95% confidence?

15.39 Neural mechanism of Valium. Valium (diazepam) is a widely prescribed antidepressant and sedative. A study investigated how Valium works by comparing its effect on sleep in 7 knock-in mice, in which the $\alpha 2GABAA$$\alpha_2 \mathrm{GABA}_A$ receptor is insensitive to Valium, and in 8 wild-type control mice. The study found that Valium reduced sleep latency in both groups, with no significant difference between the groups. According to the study authors, this lack of significance “is related to the large inter-individual variability that is also reflected in the low power (20%) of the test.”[17]

a. Explain in simple language why tests having low power often fail to give evidence against a null hypothesis even when the null hypothesis is really false.

b. Which aspects of this experiment most likely contributed to the low test power?

15.40 Treating knee pain. Arthroscopic surgery is a minimally invasive surgical procedure on a joint, most commonly performed for persistent knee pain. Several uncontrolled, retrospective studies of the procedure reported pain relief among patients following the surgery.

a. What can you conclude from these studies? What do they say about the benefits of arthroscopic surgery?

b. A recent double-blind, randomized, placebo-controlled experiment found that outcomes after arthroscopic surgery are no better than after a sham (placebo) procedure.[18] How do you interpret this result? What does it say about the benefits of arthroscopic surgery?

c. The controlled experiment was “designed to have 90 percent power, with a two-sided type I error of 0.04, to detect a moderate effect size (0.55) between the placebo group and the combined arthroscopic-treatment groups.” How is this discussion of power important in interpreting the results of the controlled experiment? What do you conclude about the benefits of arthroscopic surgery in treating knee pain?

15.41 Detecting acid rain: power. Exercise 15.7 concerned detecting acid rain (rainfall with pH less than 5) from measurements made on a sample of n days when considering different sample sizes n. That exercise shows how the *P*-value for an observed sample mean $\bar{x}$ changes with n. It would be wise to do power calculations before deciding on a sample size. Suppose that pH measurements follow a Normal distribution with standard deviation $\sigma = 0.5$. You plan to test the hypotheses

$$H_0: \mu = 5 \quad H_a: \mu < 5$$

$$\begin{aligned} H_0 &: \mu = 5 \\ H_a &: \mu < 5 \end{aligned}$$

at the 5% level of significance. You want to use a test that will almost always reject H_0 when the true mean pH is 4.7. Use the *Power of a Test* applet to find the power against the alternative $\mu = 4.7$ for samples of size $n = 5$, $n = 15$, and $n = 40$. What happens to the power as the size of the sample increases? Which of these sample sizes are adequate for use in this setting?

15.42 Detecting acid rain: power by hand. Even though software is used in practice to calculate power, doing the work by hand in a few examples will build your understanding of this concept. Find the power of the test in Exercise 15.41 for a sample of size $n = 15$ by following these steps.

a. Write the z test statistic for a sample of size 15. What values of z lead to rejecting H_0 at the 5% significance level?

b. Starting from your result in part a, what values of x-$\bar{x}$ lead to rejecting H0H_0?

c. What is the probability of rejecting H0H_0 when μ=4.7?$\mu = 4.7?$ This probability is the power against this alternative.

15.43 Power of a two-sided test. The National Center for Health Statistics reports that the systolic blood pressure for males 35 to 44 years of age has mean 128 mm Hg and standard deviation 15 mm Hg. The medical director of a large company looks at the medical records of 72 executives in this age group and finds that the mean systolic blood pressure in this sample is x-=126.07$\bar{x} = 126.07$ mm Hg. At a 5% significance level, the data fail to show significant evidence that the mean blood pressure of a population of executives differs from the national mean μ=128$\mu = 128$ mm Hg. The medical director now wonders if the test used would detect an important difference if one were present.

a. State the null and alternative hypotheses in mathematical form.

b. Use the *Statistical Power* applet to find the power of the two-sided test against the alternative μ=134$\mu = 134$ mm Hg.

APPLET

c. Use the applet to calculate the power of the test against the alternative μ=122$\mu = 122$ mm Hg. Can the test be relied on to detect a mean that differs from 128 mm Hg by 6 mm Hg?

d. If the alternative were farther from H0,H_0, say, μ=136$\mu = 136$ mm Hg, would the power be higher or lower than the values calculated in parts b and c?

CHAPTER 16 From Chance to Inference: Part II Review

Clarence Holmes/age fotostock/SuperStock

IN THIS CHAPTER WE COVER...

- Part II Summary
- Comprehensive Review Exercises
- Advanced Topics (Optional Material)
- Online Data Sources
- EESEE Case Studies

We collect data about what we do not know because of variability and uncertainty in nearly all aspects of life. That is, the inability to know for certain what the next observation will be is the reason we study random events and the **laws of probability** that govern them. The study of randomness has a number of applications.

- We use probability models to describe the distribution of a random variable and to assess how likely a particular outcome, or a range of outcomes, may be.
- We use various probability rules to compute the probability of more complex events or even to refine a probability in light of additional information (conditional probability).
- We also use chance to reduce bias when collecting sample data and we build probability models to quantify the uncertainty inherent to the sampling process.

Probability is also at the core of **statistical inference**. We draw conclusions about populations on the basis of random samples and use probability to indicate how reliable the conclusions are. A confidence interval estimates an unknown parameter. A hypothesis test shows how strong the evidence is for refuting a claim about a parameter.

The probabilities in both confidence intervals and hypothesis tests tell us what would happen if we used the method for the interval or test very many times.

- A confidence level is the success rate of the method for a confidence interval. This is the probability that the method actually produces an interval that contains the unknown parameter. A 95% confidence interval gives a correct result 95% of the time when we use it repeatedly. That is, when we produce *one* 95% confidence interval, we do not know whether it gave a correct result, but we are 95% confident that it did.
- A *P*-value tells us how surprising the observed outcome would be if the null hypothesis was true. That is, *P* is the probability that the test would produce a result at least as extreme as the observed result if the null hypothesis really was true. Very surprising outcomes (small *P*-values) are good evidence that the null hypothesis is not true.

Figures 16.1 and 16.2 use the z procedures introduced in Chapter 14 to present in picture form the big ideas of confidence intervals and hypothesis tests. These ideas are the foundation for the rest of this book. We will have much to say about many statistical methods and their use in practice. In every case, the basic reasoning of confidence intervals and hypothesis tests remains the same.

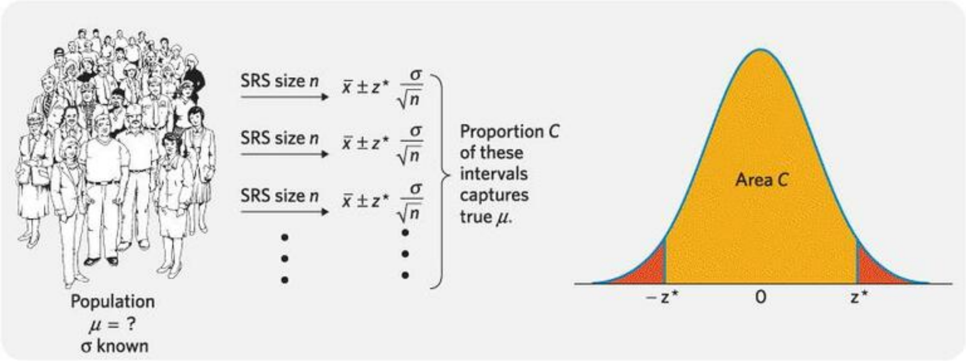

SRS size n
$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
SRS size n
$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
SRS size n
$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
Proportion C of these intervals captures true μ.
Area C
$-z^*$
0
z^*
Population
$\mu = ?$
σ known

Figure 16.1

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 16.1 Statistics in summary: The Idea of a confidence interval.

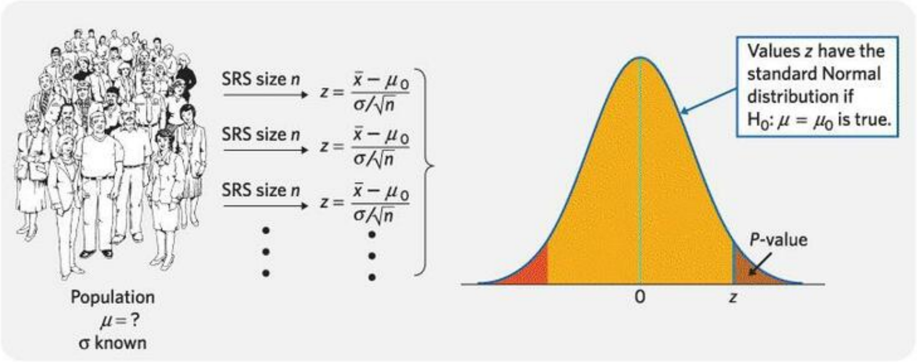

SRS size n
$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
SRS size n
$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
SRS size n
$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
Values z have the standard Normal distribution if H_0: $\mu = \mu_0$ is true.
P-value
0
z
Population
$\mu = ?$
σ known

Figure 16.2

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 16.2 Statistics in summary: The idea of a hypothesis test.

PART II SUMMARY

Here are the most important skills you should have acquired from reading Chapters 9 through 15.

A. Probability

1. Recognize that some phenomena are random. Probability describes the long-run regularity of random phenomena.
2. Understand that the probability of an event is the proportion of times the event occurs in very many repetitions of a random phenomenon. Use the idea of probability as long-run proportion to think about probability.
3. Identify the sample space for the random event, and determine whether it is discrete or continuous.
4. Use basic probability rules to detect illegitimate assignments of probability: Any probability must be a number between 0 and 1, and the total probability assigned to all possible outcomes must be 1, the probability of the entire sample space.
5. Use basic probability rules to find the probabilities of events that are formed from other events. The probability that an event does not occur is 1 minus its probability of occurring. If two events are disjoint, the probability that one or the other occurs is the sum of their individual probabilities.
6. Find probabilities in a discrete probability model by adding the probabilities of individual, disjoint outcomes. Find probabilities in a continuous probability model as areas under a density curve.
7. Use the notation of random variables to make compact statements about random outcomes, such as P(x-≤4)=0.3.$P(\bar{x} \le 4) = 0.3.$ Be able to interpret such statements.
8. Understand that the mean μμ and the standard deviation σσ of a random variable are two important summary values of its probability distribution.

B. General Rules of Probability (Optional)

1. Use Venn diagrams to picture relationships among several events.
2. Use the general addition rule to find probabilities that involve overlapping events.
3. Understand the idea of independence. Judge when it is reasonable to assume independence as part of a probability model. Don't assume that events are independent without good reason.
4. Use the multiplication rule for independent events to find the probability that all of several independent events occur.
5. Understand the idea of conditional probability—for instance, as the probability of some outcome in a subset of the sample space. Find conditional probabilities for individuals or outcomes, chosen at random, from a table of counts of possible outcomes.
6. Use the general multiplication rule to find the joint probability of A and B, $P(A \text{ and } B)$.
7. Use two-way tables or tree diagrams to organize several-stage probability models.
8. Understand the idea of Bayes's theorem. Use the formula or a tree diagram to find a conditional probability based on other, known probabilities.

C. Density Curves and Normal Distributions

1. Know that density curves represent probability distributions for continuous variables. The total area under a density curve is 1, the probability of the entire sample space.
2. Know that any given area under a density curve represents both the proportion of all corresponding observations in a population and the probability of randomly selecting one such observation from the population.
3. Recognize the shape of Normal curves and estimate by eye both the mean and the standard deviation from such a curve.
4. Use the 68–95–99.7 rule and symmetry to state what percent of the observations from a Normal distribution fall between two points when both points lie at one, two, or three standard deviations on

either side of the mean.

5. Find the standardized value (zz-score) of an observation. Interpret z z-scores and understand that any Normal distribution becomes standard Normal N(0,1)$N(0,1)$ when standardized.
6. When a variable has a Normal distribution with a stated mean μμ and standard deviation σ,$\sigma,$ obtain the proportion (probability) of values above a stated number, below a stated number, or between two stated numbers.
7. When a variable has a Normal distribution with a stated mean μμ and standard deviation σ,$\sigma,$ find the point that splits the Normal distribution in two stated, complementary areas.

D. Discrete Distributions (Optional)

1. Recognize the binomial setting: a fixed number nn of independent success/failure trials with the same probability *p* of success on each trial.
2. Recognize and use the binomial distribution of the count of successes in a binomial setting.
3. Use the binomial probability formula or software to find probabilities of events involving the count XX of successes in a binomial setting for small values of n.$n.$
4. Find the mean and standard deviation of a binomial count X.$X.$
5. Recognize when you can use the Normal approximation to a binomial distribution. Use the Normal approximation to calculate probabilities that concern a binomial count X.$X.$
6. Recognize and use the Poisson distribution of the count of occurrences of a defined event in fixed, finite intervals of time or space.
7. Identify the mean and standard deviation of a Poisson distribution.
8. Use the Poisson probability formula or software to find probabilities of events involving the count XX of occurrences over a finite interval.

E. Sampling Distributions

1. Identify parameters and statistics in a study report.
2. Recognize the fact of sampling variability: A statistic will take different values when you repeat a sample or experiment.
3. Interpret a sampling distribution as describing the values taken by a statistic in all possible repetitions of a sample or experiment under the same conditions.
4. Interpret the sampling distribution of a statistic as describing the probability distribution of its possible values.

F. The Sampling Distribution of a Sample Mean

1. Recognize when a problem involves the mean x-$\bar{x}$ of a sample. Understand that x-$\bar{x}$ estimates the mean μμ of the population from which the sample is drawn.
2. Identify the mean and standard deviation of the sampling distribution of the sample mean x-$\bar{x}$ for SRSs of size nn when the mean μμ and standard deviation σσ of the population are known.
3. Understand that x-$\bar{x}$ is an unbiased estimator of μμ and that the variability of x-$\bar{x}$ about its mean μμ is smaller for larger sample sizes.
4. Understand that the sampling distribution of x-$\bar{x}$ is approximately Normal when the sample is large (central limit theorem). Use this Normal distribution to obtain probabilities that concern x-.$\bar{x}.$
5. (Optional) Use the law of large numbers to describe the behavior of x-$\bar{x}$ as the size of the sample increases.

G. The Sampling Distribution of a Sample Proportion

1. Recognize when a problem involves the proportion p^$\hat{p}$ of successes in a sample. Understand that p^$\hat{p}$ estimates the proportion *p* of successes in the population from which the sample is drawn.
2. Identify the mean and standard deviation of the sampling distribution of the sample proportion p^$\hat{p}$ for SRSs of size nn when the population proportion *p* is known.
3. Understand that p^$\hat{p}$ is an unbiased estimator of *p* and that the variability of p^$\hat{p}$ about its mean *p* is smaller for larger sample sizes.

4. Understand that the sampling distribution of p^$\hat{p}$ is approximately Normal when the sample is large. Use this Normal distribution to obtain probabilities that concern p^.$\hat{p}$.
5. (Optional) Use the law of large numbers to describe the behavior of p^$\hat{p}$ as the size of the sample increases.

H. Confidence Intervals

1. State in nontechnical language what is meant by “95% confidence” or other statements of confidence in statistical reports.
2. Know the four-step process (page 354) for any confidence interval.
3. Obtain a confidence interval for the mean μμ of a Normal population with known standard deviation σ.σ.
4. Understand how the margin of error of a confidence interval changes with the sample size and the level of confidence *C*.
5. (Optional) Find the sample size required to obtain a confidence interval of specified margin of error mm when the confidence level and other information are given.
6. Identify sources of error in a study that are *not* included in the margin of error of a confidence interval, such as undercoverage or nonresponse.

I. Hypothesis Tests

1. State the null (H0H_0) and alternative (Ha)(H_a) hypotheses in a testing situation when the parameter in question is a population mean μ.μ.
2. Explain in nontechnical language the meaning of the *P*-value when you are given the numerical value of *P* for a test.
3. Know the four-step process (page 365) for any hypothesis test.
4. Obtain the one-sample zz test statistic and the *P*-value for both one-sided and two-sided tests about the mean μμ of a Normal population with known standard deviation σ.σ.
5. Assess the strength of the evidence against H0H_0 with the *P*-value, understanding that the smaller a *P*-value is, the more evidence it provides against H0H_0 (that is, the more statistically significant the

findings are).

6. Recognize that when the *P*-value isn't small, the evidence against H_0 is inconclusive and you can only fail to reject H_0.
7. Know how to compare the *P*-value, if necessary, with a fixed significance level α to decide whether *P* is small enough to reject H_0.
8. Recognize that hypothesis testing does not measure the size or importance of an effect. Explain why a small effect can be statistically significant in a large sample and why a large effect can fail to be statistically significant in a small sample.
9. (Optional) Recognize that the decision to reject or fail to reject H_0 may be the wrong decision. Identify the two types of error that can be made: A Type I error occurs if we reject H_0 when it is actually true, and a Type II error occurs if we fail to reject H_0 when actually H_a is true.
10. (Optional) Find the probability of making a Type I error for a test with a fixed significance level α. Obtain the probability of making a Type II error for a given alternative value of the parameter. Know that the power of a test against a specific alternative is 1 minus the probability of a Type II error.

J. Conditions for Inference

1. Recognize that any inference procedure acts as if the data were properly produced for this particular procedure.
2. Identify potential problems in the data collection process that would introduce systematic bias. The z confidence interval and test require that the data be an SRS from the population of interest.
3. Plot the data to check for deviations from Normality and outliers. The z confidence interval and test are robust to deviations from Normality as long as the sample size is large enough to compensate.
4. Recognize whether the population standard deviation σ is known. The z confidence interval and test require σ to be known. This is not a common situation. [When σ is not given, a *t* confidence interval or test may be performed instead. See Chapter 17.]

COMPREHENSIVE REVIEW EXERCISES

Review exercises help you integrate concepts and approaches presented in previous chapters, with a focus on Chapters 9 through 15.

Use the following information for Exercises 16.1 to 16.4. The 2015 National Vital Statistics Report provides information on how the nearly 4 million live births that occurred in 2014 broke down by birth order. Choose at random a baby born in the United States in 2014:[1]

Baby's birth order	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Probability	*0.391*	*0.319*	*0.168*	*0.072*	*0.028*	*0.012*	*0.005*	*0.005*

(The few babies who were 9th or more in birth order are included in the "8" birth order group.)

16.1 Birth order: probabilities. Call XX the birth order of a randomly selected baby born in the United States in 2014.

a. Check that the distribution of XX satisfies the two requirements for a legitimate discrete probability model.
b. Describe in words the events P(X≤2)$P(X \le 2)$ and P(X>5).$P(X > 5)$. What is the probability of these events?
c. Write the event "a woman gives birth in 2014 to her third child" in terms of X.X. What is the probability of this event?

16.2 Birth order: random variable. Call XX the birth order of a randomly selected baby born in the United States in 2014.

a. Explain briefly what makes XX a random variable.
b. Ignoring the fact that a few births were of the 9th or greater order, obtain the mean of X.X. Interpret this value in context.
c. The standard deviation of XX is σ=1.259.$\sigma = 1.259$. Interpret this value in context.

16.3 Birth order: average. Consider the average birth order, x-,$\bar{x}$, in a random sample of 100 babies born in the United States in 2014.

a. What are the mean and the standard deviation of the sampling distribution of x-$\bar{x}$ in random samples of n=100?$n = 100?$
b. What is the shape of the sampling distribution of x-$\bar{x}$ in random samples of n=100?$n = 100?$
c. What is the probability that one random sample of 100 U.S. births in 2014 would have a mean birth order lower than 2.0?

16.4 Birth order: first-time mothers. In 2014, 39.1% of the nearly 4 million live births in the United States were to first-time mothers.

a. Is this value a statistic or a parameter? Explain your answer.
b. What are the properties of the sampling distribution of the proportion of births to first-time mothers in random samples of 500 births from 2014?
c. If we took a random sample of 500 birth certificates from 2014, what is the probability that the sample would contain less than 35% of first births?

16.5 Accidents on a bike path. Examining the location of accidents on a level, 3-mile bike path shows that they occur uniformly along the length of the path. Figure 16.3 displays the density curve that describes the distribution of accidents.

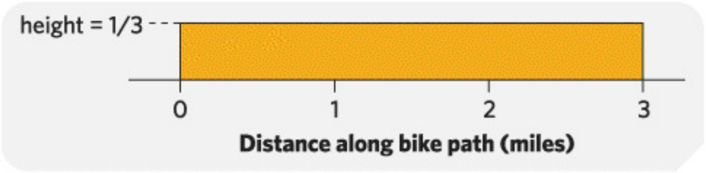


Figure 16.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 16.3 Density curve for the location of accidents along a 3-mile bike path.

a. Explain why this curve satisfies the two requirements for a density curve.
b. The proportion of accidents that occur in the middle mile of the path is the area under the density curve between 1 and 2 miles. What is this area?

c. What is the probability that an accident occurs outside the last mile?
d. What is the probability that an accident occurs either in the first half-mile or the last half-mile?

16.6 Accidents on a bike path, continued. Exercise 16.5 described the probability distribution for the location Xx of accidents on a 3-mile bike path. The probability distribution of Xx is uniform (as shown in Figure 16.3), with mean μ=1.5$\mu = 1.5$ miles and standard deviation σ=0.866$\sigma = 0.866$ mile.

a. What are the mean and the standard deviation of the sampling distribution of the mean accident location in random samples of 25 accidents?
b. Explain how we know that the sampling distribution of mean accident location in random samples of 25 accidents is approximately Normal.
c. What is the probability that the average location of a random sample of 25 accidents will be between 1 and 2 miles?
d. Individual accidents can occur at any location between 0 and 3 miles, yet nearly all average locations of random samples of 25 accidents will be between 1 and 2 miles. Explain in nontechnical language how this can be.

Use the following information for Exercises 16.7 to 16.9. Many biological measurements on the same species follow a Normal distribution quite closely. The weights of seeds of a variety of winged bean are approximately Normal with mean 525 milligrams (mg) and standard deviation 110 mg.

16.7 Winged-bean seeds. Use technology (or Table B) to answer the following questions.

a. What percent of seeds weigh more than 500 mg?
b. If we discard the lightest 10% of all seeds, what is the smallest weight among the remaining seeds?

16.8 Winged-bean seeds: individuals versus means. Use the 68–95–99.7 rule to answer the following questions.

a. You take one SRS of 100 winged-bean seeds. What approximate *range of individual weights x* do you expect to see in your sample (consider the middle 99.7%)?
b. You look at many SRSs of size 100 from the population of winged-bean seeds. What approximate *range of sample mean weights* x-$\bar{x}$ do you expect to see (consider the middle 99.7%)?

16.9 Winged-bean seeds: larger sample. In the setting of Exercise 16.8, how many winged-bean seeds must you sample to cut the range of values of x-$\bar{x}$ in half? (This is the same approach used to cut the margin of error of a confidence interval for μμ in half.) Do you expect the range of individual weights xx in the new sample to also be much narrower than in a sample of size 100? Why?

Use the following information for Exercises 16.10 to 16.12. The distribution of blood cholesterol level in the population of young men aged 20 to 34 years is close to Normal, with standard deviation σ=$\sigma =$ 41 milligrams per deciliter (mg/dl). We measure the blood cholesterol of a random sample of 14 male cross-country runners and find that they have a mean cholesterol level of 172 mg/dl. Assume that σ is the same in this population as in the general population of young adult males.

16.10 Blood cholesterol: estimation. Give a 90% confidence interval for the mean cholesterol level μμ among male cross-country runners based on our sample.

16.11 Blood cholesterol: testing. The mean blood cholesterol level for all men aged 20 to 34 years is considered to be μ=188 mg/dl.$\mu = 188\text{ mg/dl.}$ We suspect that the mean for male cross-country runners is lower.

a. State the hypotheses. Then find the test statistic and give the *P*-value based on our sample. Is the result significant at the α=0.10$\alpha = 0.10$ level? At α=0.05?$\alpha = 0.05?$ At α=0.01?$\alpha = 0.01?$
b. Explain how you can use the confidence interval you computed in Exercise 16.10 to test these hypotheses. What significance level does that 90% confidence interval yield when testing the hypotheses you stated in part a? Remember that your test was one-sided.

16.12 Blood cholesterol: inference conditions. Male cross-country runners as a group are probably more fit than young adult males as a whole, so they may also be more similar to each other physiologically. The inference procedures you used for Exercises 16.10 and 16.11 rely on a number of assumptions. Which of these assumptions may not be reasonable in this specific context? Explain your reasoning.

Use the following information for Exercises 16.13 to 16.15. Infants weighing less than 1500 grams at birth are classed as "very low birth weight." Low birth weight carries many risks. IQ tests are designed to produce IQ scores varying Normally with standard deviation $\sigma=15$.$\sigma = 15$. *They are also scaled so that the mean score in a large population should be* $\mu=100$.$\mu = 100$.

16.13 Birth weight and IQ: estimation. One study followed 113 male infants with very low birth weight to adulthood. At age 20, the mean IQ score for these men was x-=87.6.$\bar{x} = 87.6$.[2] Give a 95% confidence interval for the mean IQ score at age 20 for all very-low-birth-weight males. Use the four-step process for confidence intervals (page 354) as a guide.

16.14 Birth weight and IQ: testing. We suspect that the very-low-birth-weight population has a mean IQ score less than 100. Does the study described in Exercise 16.13 give good evidence that this is true? Use the four-step process for hypothesis tests (page 365) as a guide.

16.15 Birth weight and IQ: causation? Studies of newborns often examine the sociodemographics of their mothers as well. Such studies show that very-low-birth-weight babies are, for instance, more likely to be born to unmarried mothers and to mothers who did not complete high school.

a. Explain why this study is not an experiment.
b. Explain clearly why confounding prevents us from concluding that very low birth weight in itself reduces adult IQ.

16.16 Birth weight and IQ: error types. Exercise 16.14 asks for a hypothesis test of the null hypothesis that the mean IQ of very-low-birth-weight male babies is 100 against the alternative hypothesis that the mean is less than 100. State in words what it means to make a Type I error and a Type II error in this setting.

16.17 ASA statement on statistical significance and *P*-values. In 2016, the American Statistical Association (ASA) released an official statement on statistical significance and *P*-values addressing common misconceptions about hypothesis tests. Here are the six principles about *P*-values cited in the ASA publication:[3]

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *P*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *P*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis.

Each item in the publication (pages 131 and 132) is followed by a short, one-paragraph explanation. Which of these principles address a common misconception about the mathematical meaning of a *P*-value? Which ones address the failure to consider issues of data collection and interpretation? Which ones address both concerns? Explain your answer.

ADVANCED TOPICS (OPTIONAL MATERIAL)

Review exercises on material marked optional (*) in Chapters 9 through 15.

16.18 Freckles and sun exposure. Researchers in Germany examined the link between pigmentation (hair color, eye color, freckles) and reaction to sun exposure. After examining the reaction to midday summer sun exposure for thousands of Caucasian children, they found the following distribution in their sample:[4]

Reaction type to sun	No freckles	Freckles
(I) Always reddens, never tans	79	73
(II) Always reddens, slight tan	581	367
(III) Sometimes reddens, always tans	1025	324
(IV) Never reddens, always tans	1022	135

a. Draw a tree diagram to show the possible outcomes for a Caucasian child in this study. Let’s use the results of this study to describe a randomly chosen Caucasian child in Germany.

b. What is the probability that a child has freckles? What is the probability that a child has a type I reaction to sun exposure (always reddens, never tans)?

c. What is the conditional probability that a child has a type I reaction to sun exposure, knowing that this child has freckles? What is the conditional probability that a child has freckles, knowing that this child has a type I reaction to sun exposure?

Use the following information for Exercises 16.19 to 16.21. A large Canadian survey asked high school children, “How many days last week did you do an intense physical activity?” Call the response X for short. Based on the findings, here are separate probability models for the answers from girls and from boys:[5]

Number of days (XX)	0	1	2	3	4	5	6	7

Probability for girls	0.09	0.15	0.19	0.21	0.13	0.13	0.05	0.05
Probability for boys	0.08	0.09	0.14	0.17	0.15	0.16	0.08	0.13

16.19 Exercising in high school: association.

a. Verify that these are two legitimate discrete probability models.
b. Describe the event $X<3$$X < 3$ in words. What is $P(X<3|\text{ girl})?$ $P(X < 3|\text{girl})?$ What is $P(X<3|\text{ boy})?$$P(X < 3|\text{boy})?$
c. Express the event “worked out at least one day” in terms of $X.$$X.$ What is the probability of this event for a randomly selected boy?
d. Without doing any calculations, determine whether $X$$X$ and gender are independent in this study.

16.20 Exercising in high school: distributions.

a. Obtain the mean of $X$$X$ for girls and the mean of $X$$X$ for boys. Interpret these values in context.
b. The standard deviation of $X$$X$ is $\sigma=1.87$$\sigma = 1.87$ for girls and $\sigma=2.09$ $\sigma = 2.09$ for boys. Use this information and your computations in part a to contrast the distribution of $X$$X$ for girls and for boys.
c. In the Canadian high school population, what is the standardized score of a girl who exercises 5 days a week? Of a boy who exercises 5 days a week? What do these values reveal?

16.21 Exercising every day in high school. Consider random samples from the population of Canadian high school students.

a. What is the probability that 2 of a random sample of 10 girls would exercise every day of the week? What is the probability that 2 of a random sample of 10 boys would exercise every day of the week?
b. What is the probability that at least 1 girl in a random sample of 10 girls would exercise every day of the week? What is the probability that at least 1 boy in a random sample of 10 boys would exercise every day of the week?
c. How many girls would we expect, on average, to exercise every day of the week in a random sample of 10 girls? How many boys would we expect, on average, to exercise every day of the week in a random

sample of 10 boys?

d. The answers to part c may not be whole numbers. Explain briefly why this does not imply a mistake.

16.22 Smoking and social class. As the dangers of smoking have become more widely known, clear class differences in smoking have emerged. British government statistics classify adult men by occupation as “managerial and professional” (30% of the population), “intermediate” (29%), or “routine and manual” (34%). A survey finds that 20% of men in managerial and professional occupations smoke, 29% of men in the intermediate group smoke, and 38% of men in routine and manual occupations smoke.[6]

a. Make a tree diagram that displays the percent of all adult British men who smoke.
b. Write each of the percents provided here, using probability notations such as *P*(event).

16.23 Comorbidity. The U.S. National Institutes of Health explains on its website that the term *comorbidity* describes two or more disorders or illnesses occurring in the same person. This concept is illustrated with the statement that “compared with the general population, people addicted to drugs are roughly twice as likely to suffer from mood and anxiety disorders.”[7] Consider the general American population. In this sample space, use the notation DA for the event “has a drug addiction” and MAD for the event “has a mood and anxiety disorder.”

a. Express the quoted statement using probability notations such as *P*(event).
b. Based on the information provided, are the events DA and MAD disjoint? Are they independent? Explain briefly.

16.24 Cystic fibrosis. Cystic fibrosis (CF) is a lung disorder that often results in death. It is inherited, but it can be inherited only if both parents are carriers of an abnormal gene. In 1989, the CF gene that is abnormal in carriers of cystic fibrosis was identified. The probability that a randomly chosen person of European ancestry carries an abnormal CF gene is 1/25.

(The probability is less in other ethnic groups.) The CF20m test detects most, but not all, harmful mutations of the CF gene. This test is positive for 90% of people who are carriers. It is (ignoring human error) never positive for people who are not carriers. What is the probability that a randomly chosen person of European ancestry gets a positive result from the CF20m test?

16.25 Family planning. Many parents wish to have at least one boy and one girl. For simplicity, let’s ignore twins and other multiple births and assume that successive births are independent events with a 50-50 chance that the baby is a boy or a girl.

a. What is the distribution of the number of boys in a family with 5 children?
b. What is the probability that a family with 5 children has 5 boys? What is the probability that a family with 5 children has 4 boys and 1 girl?
c. What is the conditional probability that a couple will have a baby girl next if they already have four boys? Explain how and why this probability is different from the probabilities you calculated in part b.

16.26 More on winged-bean seeds. In Exercise 16.7 you examined the distribution of seed weights for a variety of winged bean. This distribution is approximately Normal, with mean $\mu=525$$\mu = 525$ mg and standard deviation $\sigma=110$ mg.$\sigma = 110$ mg. We consider selecting at random 4 seeds of this variety.

a. What is the probability that any one seed weighs more than 525 mg? Between 500 and 550 mg?
b. What is the probability that all 4 seeds in the sample weigh more than 525 mg? What is the probability that the average weight of the 4 seeds is greater than 525 mg? Which of these two probabilities is the smallest? Explain in nontechnical terms why it is so.
c. What is the probability that all 4 seeds in the sample weigh between 500 and 550 mg? What is the probability that the average weight of the 4 seeds is between 500 and 550 mg? How does this relate to your explanation for part b?

The next two exercises guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. They cover the concept of Normal quantile plots from Chapter 11. Note that short answers are not available for these exercises to enable their use for class work or assignments.

16.27 Estimating body fat in men. The data file *Large.Bodyfat* contains data on a number of body measurements for 252 adult men. In Exercise 8.16 (page 208) you made histograms for the distributions of the men’s body weights (in pounds) and heights (in inches). Make a Normal quantile plot for each variable and use them to complement your description of these two distributions.

16.28 Ant morphology. The file *Large.Ant* contains data on 179 adult carpenter ants (*Camponotus floridanus*) sampled from a mature ant colony. In Exercise 8.18 (page 209) you plotted the distribution of scape length (in microns, μμm), which is a simple and common proxy for body size in ants.

a. Create a Normal quantile plot for the distribution of scape lengths. Are scape lengths approximately Normally distributed in the population of adult carpenter ants?
b. The researchers determined from the observed variations in body proportions that there are two types of workers among adult carpenter ants, which they labeled major workers and minor workers. Make a separate Normal quantile plot for the distribution of scape lengths for minor and of major workers. Are scape lengths approximately Normally distributed in each subpopulation?
c. Does your analysis support the researchers’ conclusion that there are two types of workers among adult carpenter ants? Briefly explain your answer.

16.29 The first kid has higher IQ. Does the birth order of a family’s children influence their IQ scores? A careful study of 241,310 Norwegian 18- and 19-year-olds found that firstborn children scored 2.3 points higher on IQ tests, on average, than second children in the same family. This difference was highly significant (P<0.001).$(P < 0.001)$. A commentator said, “One puzzle highlighted by these latest findings is why certain other within-family studies have failed to show equally consistent results. Some of these previous null findings, which have all been obtained in much smaller samples, may be explained by inadequate statistical power.”[8]

a. Explain in simple language why tests having low power often fail to give evidence against a null hypothesis even when the hypothesis is really false.
b. A population mean IQ is typically 100. The Norwegian study found an effect size of 2.3 points. Is that a large effect size?
c. The study examined 18- and 19-year-olds. By definition, firstborn children are older than second children in the same family. What confounding variable likely plays a role in the findings of this study?

16.30 Oat bran. Does eating oat bran lower cholesterol? At a time when oat bran was something of a fad, a paper in the *New England Journal of Medicine* found that it had no significant effect on cholesterol.[9] The paper reported a study with just 20 subjects. Letters to the journal denounced publication of a conclusion of no effect from a study with very low power.

a. Explain why the nonsignificant P-value does not imply that oat bran consumption has no effect on cholesterol.
b. Explain why lack of significance in a study with low power gives reason to question the usefulness of the study publication.

ONLINE DATA SOURCES

Here are a few regularly updated, online data sources relevant to the life sciences that can help further illustrate and practice the concepts of Chapters 9 through 15.

- www.cdc.gov/mmwr/ CDC's *Morbidity and Mortality Weekly Report* (*MMWR*)
- www.cdc.gov/nchs/ National Center for Health Statistics: Click on "Data and Tools" tab
- www.cdc.gov/nchs/nvss/births.htm Birth data for the United States, National Vital Statistics System (updated every year)
- www.cdc.gov/nchs/hus "Health, United States": Comprehensive health statistics, updated every year
- www.gallup.com/topic/category_wellbeing.aspx Gallup's well-being surveys
- fivethirtyeight.com/science/ Health and science section of the data analytic website FiveThirtyEight
- www.pewresearch.org/science/ Pew Research Center, section on scientific topics
- apps.who.int/gho/data/ World Health Organization, data repository
- www.statcan.gc.ca Statistics Canada
- www.healthycanadians.gc.ca/science-research-sciences-recherches/index-eng.php Canadian government website, section on science, research, and data

EESEE CASE STUDIES

The Electronic Encyclopedia of Statistical Examples and Exercises (EESEE) is available on the companion website. These more elaborate stories, with data, provide settings for longer case studies. Here are some suggestions for EESEE stories that apply the ideas you have learned in Chapters 9 through 15.

- **Anecdotes of Probability.** Questions 3 and 4.
- **Anecdotes of Significance Testing.** All questions.
- **Checkmating and Reading Skills.** Questions 1(b), 2, and 3.
- **Habitat of the Spotted Owl.** Question 6.

PART III

Kwangshin Kim/Science Source

With the principles of statistical inference in hand, we proceed to practice—that is, to inference procedures applied to specific situations depending on data type and study design.

The first five chapters of Part III concern inference about the distribution of a single variable and inference for comparing the distributions of two variables. Chapters 17 and 18 deal with quantitative data, whereas Chapters 19 through 21 deal with categorical data.

The next three chapters concern inference about relationships among variables. In Chapter 22, both variables are categorical. Chapter 23 considers inference in the setting of linear regression when both variables are quantitative. Chapter 24 covers inference for quantitative data when comparing the mean responses in more than two groups—which collectively form a categorical explanatory variable.

These chapters introduce new computational methods without introducing fundamentally new concepts. Be sure to review the four-step process for a confidence interval (page 354) and for a hypothesis test (page 365). There is more to inference than computations: Any inference method is useful only under certain conditions that must be carefully checked, and the confidence intervals and *P*-values we obtain should always support a comprehensive conclusion in context.

Chapter 25 reviews this part of the text, provides more comprehensive exercises, and offers guidance and practice with selecting an appropriate inference procedure.

Statistical Inference

INFERENCE ABOUT QUANTITATIVE VARIABLES

Chapter 17 Inference about a Population Mean

Chapter 18 Comparing Two Means

INFERENCE ABOUT CATEGORICAL VARIABLES

Chapter 19 Inference about a Population Proportion

Chapter 20 Comparing Two Proportions

Chapter 21 The Chi-Square Test for Goodness of Fit

INFERENCE ABOUT RELATIONSHIPS

Chapter 22 The Chi-Square Test for Two-Way Tables

Chapter 23 Inference for Regression

Chapter 24 One-Way Analysis of Variance: Comparing Several Means

Chapter 25 Statistical Inference: Part III Review

CHAPTER 17 Inference about a Population Mean

wavebreakmedia/Shutterstock.com

IN THIS CHAPTER WE COVER...

- Conditions for inference
- The *t* distributions
- The one-sample *t* confidence interval
- The one-sample *t* test
- Matched pairs *t* procedures
- Robustness of *t* procedures

This chapter describes confidence intervals and hypothesis tests for the mean μ of a population. In Chapter 14 we used the z procedures in this setting to introduce the ideas of confidence intervals and hypothesis tests. The z procedures for μ require that we know the value of the population standard deviation σ, which is only rarely the case. We now consider the much more common situation in which σ is unknown. Fortunately, the details of confidence intervals and hypothesis tests for μ change only slightly when we don't know σ and we can interpret the results in just the same way.

CONDITIONS FOR INFERENCE

Confidence intervals and hypothesis tests for a population mean μμ require an unbiased estimate and a known sampling distribution for the estimate, so that a confidence level or a *P*-value may be computed. The methods described in this chapter are based on the sampling distribution of the sample mean x-$\bar{x}$ and require the following conditions to be satisfied.

CONDITIONS FOR INFERENCE ABOUT A MEAN

- We can regard our data as a **simple random sample** (SRS) from a much larger[1] population. This condition is very important.
- Observations from the population have a **Normal distribution** with mean μμ and standard deviation σ.σ. In practice, the inference procedures in this chapter are **robust** to deviations from the Normality condition when the sample is large enough.

When the conditions for inference are satisfied, the sample mean x-$\bar{x}$ has a Normal distribution, with mean equal to the population mean μμ and standard deviation equal to σ/n.$\sigma/\sqrt{n}$. When we don't know the population standard deviation σ,σ, we estimate it by the sample standard deviation s.s. We can then estimate the standard deviation of x-$\bar{x}$ by s/n.$s/\sqrt{n}$. This quantity is called the *standard error* of the sample mean x-,$\bar{x}$, often abbreviated in reports as "SEM" or "SE Mean."

STANDARD ERROR

When the standard deviation of a statistic is estimated from sample data, the result is called the **standard error** of the statistic. The standard error of the sample mean x-$\bar{x}$ is s/n.$s/\sqrt{n}$.

What do these values represent? The standard deviation ss of the nn individuals in a random sample is an estimate of the standard deviation σσ of all the individuals in the population from which that random sample was taken. Therefore, ss *estimates how individual observations vary.*

In contrast, the standard error s/n$s/\sqrt{n}$ of the mean x-$\bar{x}$ of nn individuals in a random sample is an estimate of the standard deviation σ/n $\sigma/\sqrt{n}$ of the sampling distribution of x-$\bar{x}$ for all random samples of nn observations taken from the same population. Therefore, s/n$s/\sqrt{n}$ *estimates how sample means vary.*

APPLY YOUR KNOWLEDGE

17.1 Neuronal activity. Neurons fire action potentials (“spikes”) at different frequencies under different circumstances. A neuron is recorded in an awake-behaving monkey looking at a luminous dot 40 centimeters away to determine the neuron’s baseline activity. Over the course of 72 trials randomly interspersed with other conditions, that neuron is found to have an average baseline activity of x-=27.8$\bar{x} = 27.8$ spikes per second with standard deviation s=32.2 $s = 32.2$ spikes per second. What is the standard error of the mean?

17.2 Spider webs. Researchers examined the mechanical properties of webs spun by the orb spider, *Nephelia edulis*.[2] A table in their published report gives summary statistics for a number of variables in the form of the mean plus or minus the standard error of the mean. This form is very common in scientific publications. For 15 adult female spiders, web areas (in square centimeters, cm^2) are 634±30.634 ± 30. What are x-$\bar{x}$ and ss for these 15 spider webs?

Sami Sarkis/Exactostock-1672/SuperStock

THE *t* DISTRIBUTIONS

As we saw in Chapter 13, the sampling distribution of x-$\bar{x}$ is the N(μ,σ/n) $N(\mu, \sigma/\sqrt{n})$ distribution when the population is Normal with mean μμ and standard deviation σ.σ. Therefore, the value

z=x--μσ/n

$$z=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

follows the standard Normal distribution N(0,1).$N(0,1)$. In the rare instances when we know the value of σ,σ, we base confidence intervals and hypothesis tests for μμ on the one-sample zz statistic (described in Chapter 14). Otherwise, we can substitute the standard error s/n$s/\sqrt{n}$ of x-$\bar{x}$ for its standard deviation σ/n.$\sigma/\sqrt{n}$. The statistic that results does not have the standard Normal distribution, but rather has a distribution that is new to us, called a *t distribution.*

THE ONE-SAMPLE *t* STATISTIC AND THE *t* DISTRIBUTIONS

Draw an SRS of size nn from a large population that has the Normal distribution with mean μμ and standard deviation σ.σ. The **one-sample** tt **statistic**

t=x--μs/n

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

has the tt **distribution** with n-1$n-1$ degrees of freedom.

The tt statistic has the same interpretation as any standardized statistic: It says how far the sample statistic (x-)$(\bar{x})$ is from its mean (μ)(μ) in standard deviation units (s/n).$(s/\sqrt{n})$. The tt distributions are a family of probability density curves with similar properties that are differentiated by a value called **degrees of freedom (df)**.

Sample standard deviations obtained with larger samples are better estimates of the unknown population standard deviation σ.σ. As a result, we need a different tt distribution for each sample size. When doing inference about a single population mean μ,$\mu,$ the tt statistic follows a tt distribution with n-1$n-1$ degrees of freedom.[3] (Other inference methods use different degrees of freedom.) We will write the tt distribution with n-1$n-1$ degrees of freedom as t(n-1)$t(n-1)$ for short.

Figure 17.1 compares the density curves of the standard Normal distribution (dotted line) and of the tt distributions with 5,10, and 20 degrees of freedom (solid lines). The figure illustrates these facts about the tt distributions:

- The density curves of the *t* distributions are similar in shape to the standard Normal curve. They are symmetric about 0, single-peaked, and “bell-shaped.”
- The spread of *t* distributions is a bit larger than that of the standard Normal distribution. The *t* distributions in Figure 17.1 have more probability in the tails and less in the center than does the standard Normal distribution. This shape reflects the fact that substituting the estimate ss for the fixed parameter σσ introduces more variation into the statistic. (Therefore, inference will be generally less precise.)
- The spread of *t* distributions also depends on the distribution’s degrees of freedom. As the degrees of freedom increase, the *t* density curve approaches the N(0,1)$N(0,1)$ curve ever more closely. Because ss estimates σσ more accurately as the sample size increases, using ss in place of σσ causes little extra variation when the sample is large.

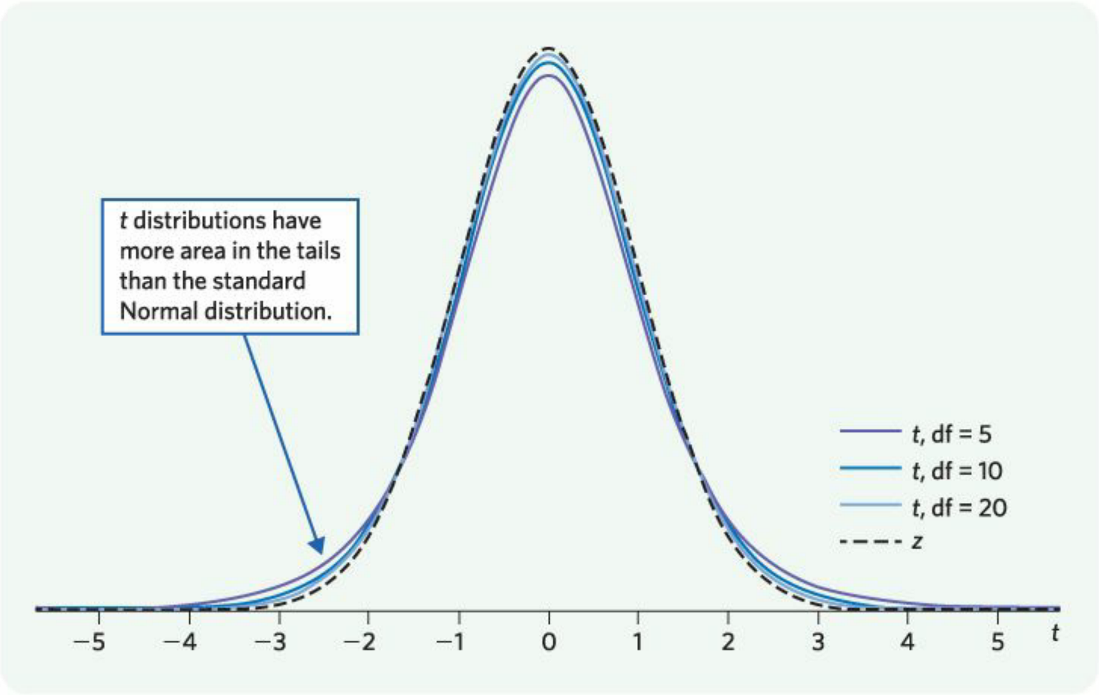


Figure 17.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.1 Density curves for the *t* distributions with 5, 10, and 20 degrees of freedom (solid lines) and for the standard Normal distribution (dotted line). All are symmetric with center 0. The *t* distributions are somewhat more spread out.

Students often wonder about the meaning of the degrees of freedom of the *t* distribution. This value helps define the exact shape of a *t* distribution. For the purpose of performing inference procedures, all you need to know is that there are many *t* distributions and you must specify, through the degrees of freedom, which *t* distribution is relevant for your computations.

Technology or printed tables like Table C in the back of the book can be used to obtain a critical value or an area under a *t* curve. The approach is similar to that covered in Chapter 11.

STATS IN YOUR WORLD

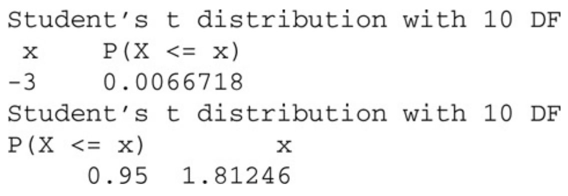
BETTER STATISTICS, BETTER BEER

The *t* distribution and the *t* inference procedures were invented by William S. Gosset (1876–1937). Gosset worked for the Guinness brewery to make better beer. He used his new *t* procedures to find the best varieties of barley and hops. Gosset's statistical work helped him become head brewer. Because of Guinness's strict nondisclosure clause, Gosset published his statistical work under the pen name “Student.” You will often see the *t* distribution called “Student's *t*” in his honor.

EXAMPLE 17.1 *t* computations[i]

Figure 17.1 shows the density curve for the *t* distribution with 10 degrees of freedom. What is the area to the left of t=-3$t = -3$ under this curve? Which point on this distribution has probability 0.05 to its right?

When using software, you need to specify the *t* distribution with its degrees of freedom and select a function that either computes a probability (area) or finds a critical value. In R, for example, the command pt(-3,10) returns the value 0.006671828 (probability, or area, to the left of t=-3)$t = -3)$ and the command qt(0.95,10) returns the value 1.812461 (critical value t*t^* with area 0.95 to its left and area 0.05 to its right). Here are the same results from Minitab, which provides a more explicit output:



```
Student’s t distribution with 10 DF
 x     P(X <= x)
-3     0.0066718
Student’s t distribution with 10 DF
P(X <= x)          x
     0.95   1.81246
```

When no technology is available, Table C can be used in much the same way. Each row in the table contains critical values for the *t* distribution whose degrees of freedom appears at the left of the row. (Because *t* distributions are symmetric about zero, and to save space, only positive *t* values are displayed.) This table provides three areas for each positive *t* value: a central area between −*t* and *t* (“Confidence level *C*”), an upper-tail area to the right of *t* (“One-sided *P*”), and a two-tailed area twice the size of the upper-tail area (“Two-sided *P*”).

To find the point on the t(10)$t(10)$ distribution with probability 0.05 to its right, look in the df 10 row and the one-sided *P* .05 column. The intersection is t*=1.812.$t^* = 1.812.$ Obtaining the area to the left of t=-3 $t = -3$ is a little more challenging. Due to space limitations in a printed table, only a dozen *t* values are printed for each distribution. Therefore,

> the best you can do is find the printed value (or values) nearest to the positive value t=3$t = 3$ in the df 10 row. There are two close values, t=2.764$t = 2.764$ and t=3.169,$t = 3.169,$ which have upper-tail areas of 0.01 and 0.005, respectively. All you can say from Table C is that the area to the left of t=-3$t = -3$ (lower-tail area) is somewhere between 0.005 and 0.01. (Technology gave us 0.007 for this area.)

The modern practice of statistics relies on technology to automate all computations. Therefore, you will likely not need to obtain critical values or areas under *t* curves as shown in Example 17.1. Instead, most users of statistics select a software function that performs all computations in the background and simply reports a confidence interval or a test *P*-value.

APPLY YOUR KNOWLEDGE

17.3 More tt computations. Use software or Table C to find

a. the critical value t*t^* corresponding to the middle 98% of the t(21)$t(21)$ distribution.

b. the critical value t*t^* with area 0.025 to the right of t*t^* under the t(24)$t(24)$ distribution.

c. the area under the t(5)$t(5)$ distribution to the right of t=3.$t = 3.$

17.4 Thinking about tt distributions. Consider the t(20)$t(20)$ and t(40) $t(40)$ distributions.

a. Which distribution is wider?

b. For the same value of t,$t,$ which distribution has the smallest tail area?

c. For the same middle area *C*, which distribution has the largest t* t^* critical value?

THE ONE-SAMPLE *t* CONFIDENCE INTERVAL

To analyze samples from Normal populations with unknown μμ and σ,σ, just replace the standard deviation of x-,$\bar{x}$, σ/n,$\sigma/\sqrt{n}$, by its standard error s/n$s/\sqrt{n}$ in the zz procedures of Chapter 14. The confidence interval and hypothesis test that result are *one-sample t procedures*. Critical values and *P*-values come from the *t* distribution with n-1$n-1$ degrees of freedom. The one-sample *t* procedures are similar in both reasoning and computational detail to the zz procedures.

THE ONE-SAMPLE *t* CONFIDENCE INTERVAL

Draw an SRS of size nn from a large population having unknown mean μμ and unknown standard deviation σ.σ. A level *C* **confidence interval for** μμ is

x-±t*sn

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t*t^* is the critical value for the t(n-1)$t\,(n-1)$ density curve with area *C* between -t*$-t^*$ and t*.t^*. Any technology with basic statistical inference capabilities will perform all necessary computations and produce the lower and upper boundaries of the confidence interval.

This interval is exact when the population distribution is Normal, and is approximately correct for large enough nn otherwise.

EXAMPLE 17.2 Testosterone and obesity in adolescent males

We follow the four-step process to obtain a confidence interval, outlined on page 354.

STATE: Obesity in adult males is associated with lower levels of sex hormone. A study investigated a possible link between obesity and plasma testosterone concentrations in adolescent males between the ages of 14 and 20 years. Here are the data for 25 obese adolescent males, measured in nanomoles per liter of blood (nmol/l):[4]

0.30 0.24 0.19 0.17 0.18 0.23 0.24 0.06 0.15
0.17 0.18 0.17 0.15 0.12 0.25 0.25 0.25 0.32
0.35 0.37 0.39 0.46 0.49 0.42 0.36

We want to estimate the mean testosterone level in obese adolescent males.

Dorling Kindersley/Getty Images

PLAN: We will estimate the mean testosterone level μμ for all obese adolescent males by giving a 95% confidence interval.

SOLVE: CHECK THE CONDITIONS FOR INFERENCE:

- We don't have an actual SRS from the population of all obese adolescent males. Scientists usually act as if a set of subjects is an SRS of the larger population of interest if there is nothing special about how the subjects were obtained. These subjects were obese young males at pubertal and postpubertal stages who were otherwise healthy. They were recruited at the Women and Children's Hospital of Buffalo, New York, the health care facility associated with the research institution conducting the study. The subjects should be representative of all obese adolescent males in the Buffalo area and perhaps the United States or countries with similar populations.
- The dotplot in Figure 17.2 does not suggest any strong departures from Normality. This is also confirmed with a Normal quantile plot (not shown).

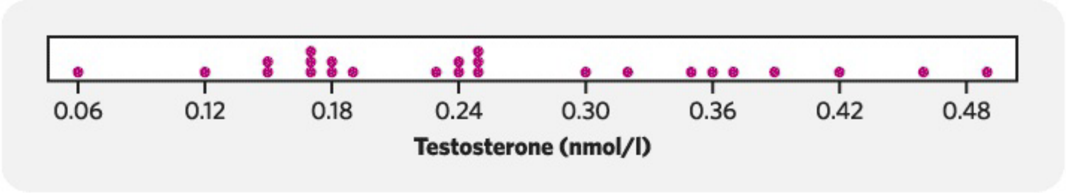
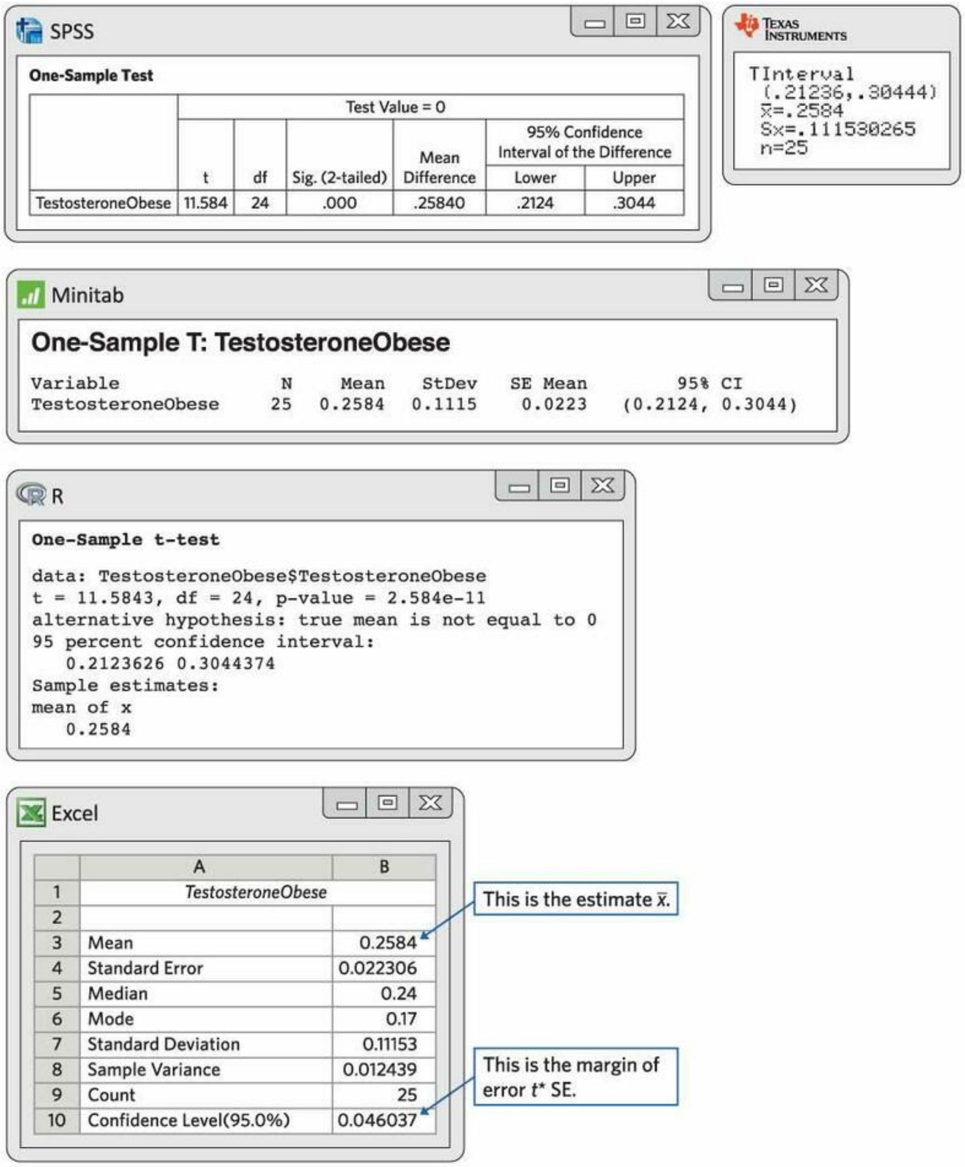


Figure 17.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.2 Dotplot of plasma testosterone levels in obese adolescent males.

SOLVE: OBTAIN THE CONFIDENCE INTERVAL: Software will find x-,$\bar{x}$, s,s, SEM, and the confidence interval from the data, as illustrated in Figure 17.3. Rounded to four decimal places, all five outputs give (0.2124, 0.3044) for the 95% confidence interval for the mean μμ of this population.



One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
TestosteroneObese	11.584	24	.000	.25840	.2124	.3044

```
TInterval
 (.21236,.30444)
 x̄=.2584
 Sx=.111530265
 n=25
```

One-Sample T: TestosteroneObese

```
Variable            N    Mean   StDev  SE Mean        95% CI
TestosteroneObese  25  0.2584  0.1115   0.0223   (0.2124, 0.3044)
```

```
One-Sample t-test

data: TestosteroneObese$TestosteroneObese
t = 11.5843, df = 24, p-value = 2.584e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
   0.2123626 0.3044374
Sample estimates:
mean of x
   0.2584
```

	A	B
1	*TestosteroneObese*	
2		
3	Mean	0.2584
4	Standard Error	0.022306
5	Median	0.24
6	Mode	0.17
7	Standard Deviation	0.11153
8	Sample Variance	0.012439
9	Count	25
10	Confidence Level(95.0%)	0.046037

This is the estimate $\bar{x}$.

This is the margin of error t^* SE.

Figure 17.3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.3 The *t* confidence interval for the testosterone data: output from a graphing calculator, three statistical programs, and a spreadsheet program.

CONCLUDE: We are 95% confident that the mean plasma testosterone level in the population of all obese adolescent males similar to the individuals studied is somewhere between 0.212 and 0.304 nmol/l. This conclusion may apply to the larger population of all obese adolescent males in the United States or elsewhere.

The computations performed by software in Example 17.2 involve the following steps:

- Summary statistics: x-=0.2584, $\bar{x} = 0.2584,$ s=0.1115, $s = 0.1115,$ and n=25 $n = 25$
- Degrees of freedom of the *t* distribution: n-1=25-1=24 $n - 1 = 25 - 1 = 24$
- Critical value for a 95% confidence level under t(24): t*=2.064 $t\,(24) :\ t^{*} = 2.064$ (rounded to three decimal places, obtained from software or Table C)
- Margin of error: m=t*s/n=2.064(0.1115/25)=0.0460 $m = t^{*}s/\sqrt{n} = 2.064\,(0.1115/\sqrt{25}) = 0.0460$
- 95% confidence interval for μ:x-±m=0.2584±0.0460, $\mu :\ \bar{x} \pm m = 0.2584 \pm 0.0460,$ or 0.2124 to 0.3044

The computations are very similar to those we performed in Chapter 14. In this case, however, because we do not know the true value of the population standard deviation σ, $\sigma,$ we use the sample standard deviation s=0.1115 $s = 0.1115$ calculated from the data. The price we pay for this approximation comes in the form of having to use the *t* critical value t*=2.064 $t^{*} = 2.064$ for 24 degrees of freedom instead of the standard Normal critical value z*=1.960. $z^{*} = 1.960.$

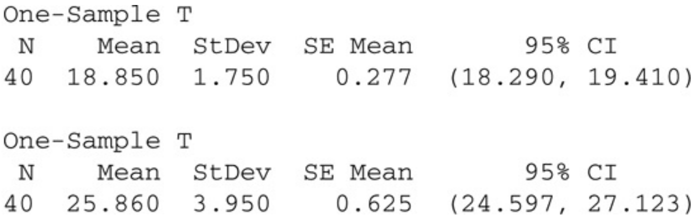
In general, software results are slightly more accurate than hand calculations, which involve some rounding at each step. The TI-83, Minitab, R, and SPSS outputs all give the estimate x- $\bar{x}$ and the confidence interval plus a selection of other information. Excel, in contrast, does not give the confidence interval. Instead, the entry labeled “Confidence Level (95.0%)” is the margin of error m, $m,$ which, as expected, is equal to half the width of the confidence interval (m=(0.3044-0.2124)/2=0.0460). $(m = (0.3044 - 0.2124)\,/2 = 0.0460).$ You can combine the mean and the margin of error to compute the confidence interval as x-±m (0.258±0.046). $\bar{x} \pm m\,(0.258 \pm 0.046).$

Because the sample in Example 17.2 was not an SRS from a well-defined population and because an observational study could have many confounding variables, we must be careful about our conclusion. For example, if testosterone levels are influenced by genetic factors, locations with a different mix of ethnicities could have a different mean μ. $\mu.$ Similarly, if testosterone levels are influenced by environmental factors, areas with different exposures also could have a different mean μ. $\mu.$ Our confidence interval definitely applies to a larger population of individuals similar to those sampled, but there is no guarantee that this larger population is all obese adolescent males in the United States or elsewhere.

APPLY YOUR KNOWLEDGE

17.5 Spider webs. Exercise 17.2 gave the summary data (mean 634, standard error 30 cm^2) for the web area of 15 female orb spiders. Assuming an approximately Normal distribution of web areas, and assuming that we can regard these 15 female orb spiders as an SRS of the population of all female orb spiders, give the 95% confidence interval for the mean web area μμ in this population.

17.6 An attractive body. Researchers recruited a sample of 40 female and 40 male heterosexual Caucasian undergraduate students from British universities for a study of the attractiveness of body types. Among adults, a body mass index (BMI, in kg/m^2) value between 18.5 and 24.9 is considered healthy, but attractiveness may not be entirely about health. Using a 3D computer avatar, participants built what they considered the ideal body of an adult of their own gender. Here is the Minitab output for the mean ideal BMI of the women (top) and of the men (bottom):[5]



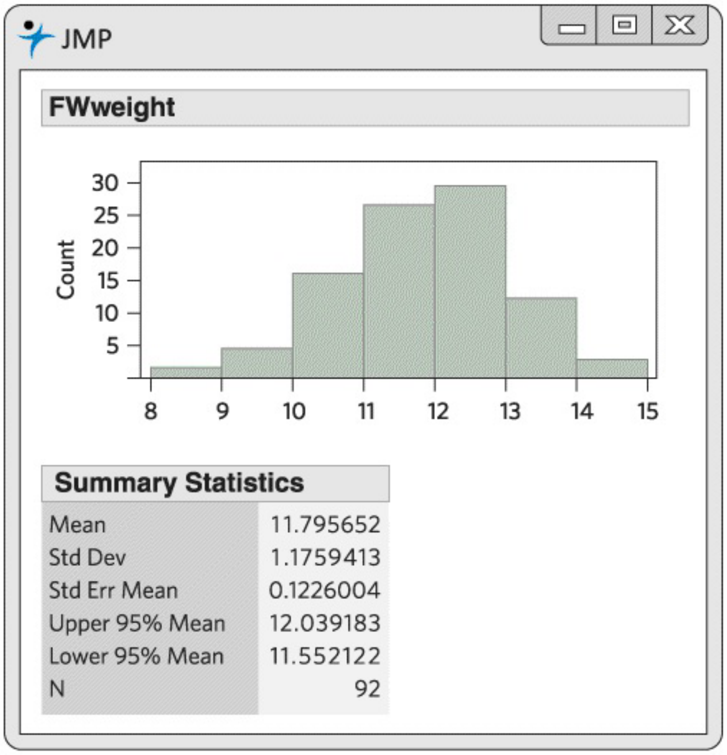
```
One-Sample T
 N     Mean   StDev  SE Mean        95% CI
40   18.850   1.750    0.277   (18.290, 19.410)

One-Sample T
 N     Mean   StDev  SE Mean        95% CI
40   25.860   3.950    0.625   (24.597, 27.123)
```

a. Interpret in context the confidence interval for the women. How does it relate to the definition of a healthy BMI?

b. Interpret in context the confidence interval for the men. How does it relate to the definition of a healthy BMI?

c. The researchers claim to answer the question of what constitutes an ideal body size. Why might the study findings not be representative of all adults in general?

17.7 Butterfly wings. Researchers studied the morphological attributes of monarch butterflies (*Danaus plexippus*), a species that undertakes large seasonal migrations over North America. They measured the forewing weight (in milligrams, mg) of a sample of 92 monarch butterflies, all of which had been reared in captivity in identical conditions.[6] Figure 17.4 shows the output from the statistical software JMP. (The data are also available in the *Large.Butterfly* the data file if you wish to practice working with your own software.) Estimate with 95% confidence the mean forewing weight of monarch butterflies reared in captivity. Follow the four-step process as illustrated in Example 17.2.



Summary Statistics

Mean	11.795652
Std Dev	1.1759413
Std Err Mean	0.1226004
Upper 95% Mean	12.039183
Lower 95% Mean	11.552122
N	92

Figure 17.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.4 Software output (JMP) for the forewing weight of monarch butterflies.

17.8 Trout habitat. A student in wildlife management studied trout habitats in the upper Shavers Fork watershed in West Virginia. The springtime water pH of 29 randomly selected tributary sample sites was found to have the following values:[7]

6.2	6.3	5.0	5.8	4.6	4.7	4.7	5.4	6.2	6.0
5.4	5.9	6.2	6.1	6.0	6.3	6.2	5.8	6.2	6.3
6.3	6.3	6.4	6.5	6.6	6.1	6.3	4.4	6.7	

Use a 90% confidence interval to estimate the mean springtime water pH of the tributary water basin around the Shavers Fork watershed. Follow the four-step process as illustrated in Example 17.2.

THE ONE-SAMPLE *t* TEST

As with confidence intervals, the *t* test is very similar to the zz test we studied in Chapter 14. We state a set of null and alternative hypotheses (H0,Ha)(H_0, H_a) about the unknown mean μμ of a population, and compute a test statistic from the sample data collected. The test *P*-value is the probability, if the null hypothesis was true, of obtaining a test statistic at least as extreme (in the direction of the alternative) as that computed. A small *P*-value provides enough evidence against H0H_0 to reject it in favor of Ha,H_a, leading us to call the results statistically significant. In contrast, a *P*-value that isn't small is inconclusive, and we can only fail to reject H0. H_0.

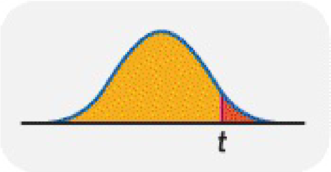
THE ONE-SAMPLE *t* TEST

Draw an SRS of size nn from a large population having unknown mean μμ and unknown standard deviation σ.σ. To **test the hypothesis** H0: μ=μ0,$\mathbf{H}_0 : \mu = \mu_0$, obtain the **one-sample *t* statistic**

t=x--μ0s/n

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

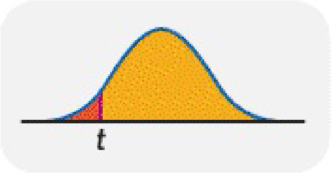
In terms of a variable TT having the t(n-1)$t\,(n-1)$ distribution, the ***P*-value** for a test of H0H_0 against



Ha:μ>μ0$H_a : \mu > \mu_0$ isis P(T≥t)$P\,(T \geq t)$

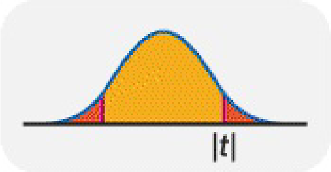
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

$H_a: \mu < \mu_0$ is $P(T \le t)$



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

$H_a: \mu \ne \mu_0$ is $2P(T \ge |t|)$



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Any technology with basic statistical inference capabilities will perform all necessary computations and provide both the test statistic and the *P*-value.

This *P*-value is exact if the population distribution is Normal; it is approximately correct for large enough n otherwise.

EXAMPLE 17.3 Children with progeria

Hutchinson-Gilford progeria syndrome is a rare genetic condition that produces rapid aging in children. In individuals with this syndrome, cardiovascular disease is a common cause of death in the teenage years. A clinical study examined the effect of treatment with the drug lonafarnib on a number of physiological outcomes. First, the researchers measured the pulse wave velocity (PWV) of 18 children diagnosed with progeria and recruited worldwide. PWV is often used as a measure of vascular stiffness, an important factor in cardiovascular health. Here are the findings (in meters per second, m/s):[8]

18.8	17.6	17.5	16.0	14.8	14.1	13.7	13.1	12.9
12.9	12.4	10.1	9.3	9.1	8.3	8.3	7.9	7.2

These data have mean x-=12.44 m/s$\bar{x} = 12.44 \text{ m/s}$ and standard deviation s=3.638 m/s.$s = 3.638 \text{ m/s}.$ In healthy children, PWV values greater than 6.6 m/s6.6 m/s would be considered abnormal. We follow the four-step process for a hypothesis test, outlined on page 365.

STATE: A study reports that the mean PWV of 18 children diagnosed with progeria is x-=12.44 m/s,$\bar{x} = 12.44 \text{ m/s},$ with standard deviation s=3.638 m/s.$s = 3.638 \text{ m/s}.$ Is this significant evidence that the mean PWV of children with progeria is abnormally high—that is, greater than 6.6 m/s?$6.6 \text{ m/s}?$

PLAN: Individual PWV values vary, so we ask the question in terms of the mean PWV μμ for the population of all children diagnosed with progeria. The null hypothesis is “no difference” from the mean μ0=6.6 m/s$\mu_0 = 6.6 \text{ m/s}$ at the high end of what is still considered normal for healthy children. The alternative is one-sided, because we want to know if, on average, PWV is abnormally high (>6.6 m/s)$(> 6.6 \text{ m/s})$ in children with progeria.

H0:μ=6.6Ha:μ>6.6

$$\begin{aligned} &H_0 : \mu = 6.6 \\ &H_a : \mu > 6.6 \end{aligned}$$

SOLVE: CHECK THE CONDITIONS FOR INFERENCE. We are willing to regard these 18 children with progeria recruited worldwide as an SRS from the population of all children with progeria. The dotplot in Figure 17.5 shows no obvious departure from Normality and no outlier. The *t* test should be accurate.

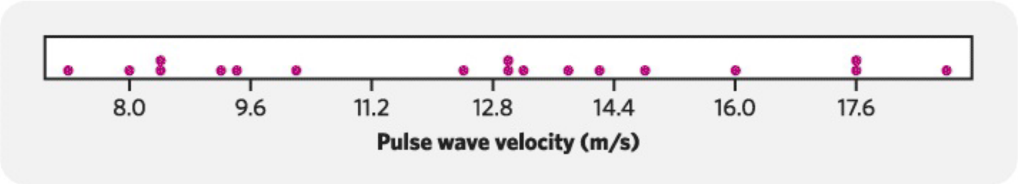


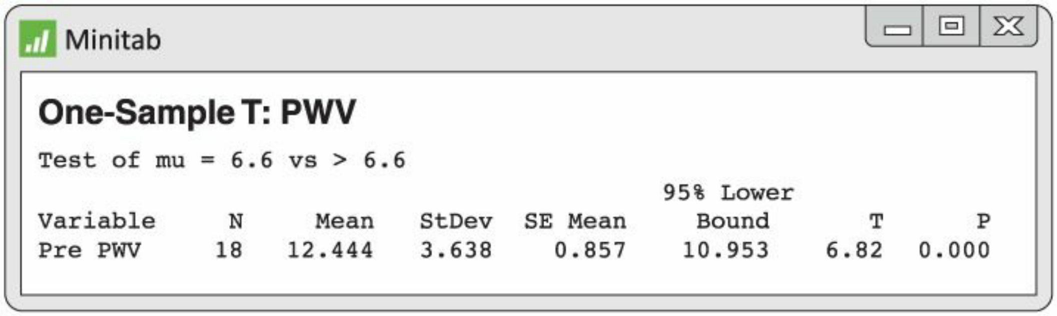
Figure 17.5

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

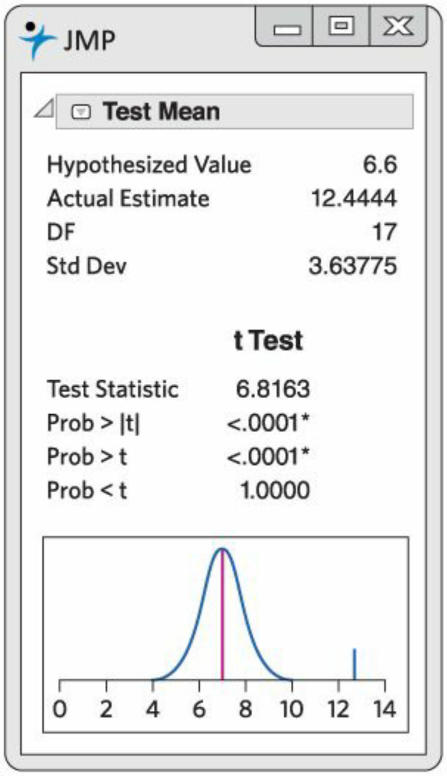
FIGURE 17.5 Dotplot of pulse wave velocity (PWV) in children with progeria.

SOLVE: OBTAIN THE TEST STATISTIC AND *P*-VALUE.

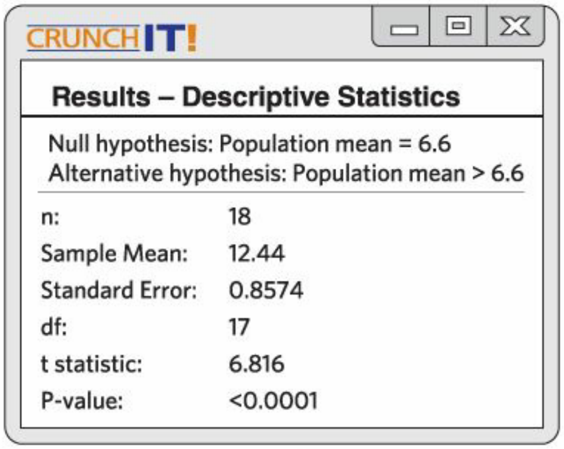
Software will find the test statistic *t* and the *P*-value from the data, as shown in Figure 17.6. All five outputs give $t=6.82$.$t = 6.82$. The test *P*-value is reported as 1.5E−6 (or, because this is such a small value, a sufficient indication such as $P<0.0001$).$P < 0.0001$).



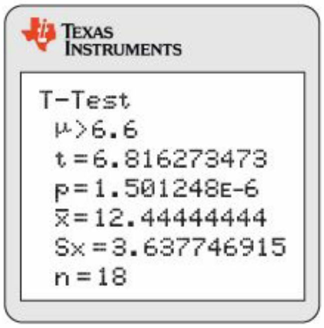
```
One-Sample T: PWV

Test of mu = 6.6 vs > 6.6

                                            95% Lower
Variable     N     Mean   StDev  SE Mean      Bound       T       P
Pre PWV     18   12.444   3.638    0.857     10.953    6.82   0.000
```



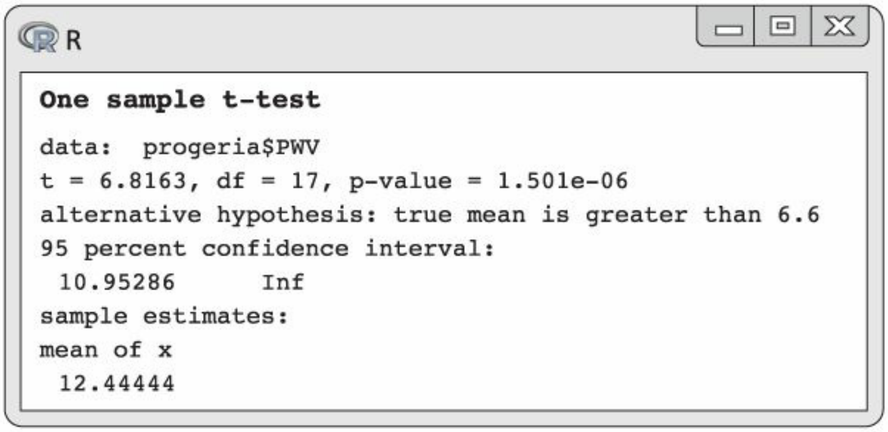
```
Hypothesized Value        6.6
Actual Estimate       12.4444
DF                         17
Std Dev               3.63775

Test Statistic     6.8163
Prob > |t|         <.0001*
Prob > t           <.0001*
Prob < t           1.0000
```



```
Results – Descriptive Statistics

Null hypothesis: Population mean = 6.6
Alternative hypothesis: Population mean > 6.6

n:                 18
Sample Mean:       12.44
Standard Error:    0.8574
df:                17
t statistic:       6.816
P-value:           <0.0001
```



```
T-Test
 μ>6.6
 t=6.816273473
 p=1.501248E-6
 x̄=12.44444444
 Sx=3.637746915
 n=18
```



```
One sample t-test

data:  progeria$PWV
t = 6.8163, df = 17, p-value = 1.501e-06
alternative hypothesis: true mean is greater than 6.6
95 percent confidence interval:
 10.95286      Inf
sample estimates:
mean of x
 12.44444
```

Figure 17.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.6 The *t* test for the study of progeria: output from a graphing calculator and four statistical programs.

> **CONCLUDE:** There is very strong evidence $(P = 0.000002)$ that the true mean PWV of all children diagnosed with progeria is greater than $6.6\ \text{m/s}$. Therefore, on average, children with progeria have abnormally high PWV, which is indicative of abnormally high vascular stiffness.

The computations performed by software in Example 13.3 involve the following steps:

- Summary statistics: $\bar{x} \approx 12.44$, $s \approx 3.638$, and $n = 18$
- Degrees of freedom of the *t* distribution: $n - 1 = 18 - 1 = 17$
- Test statistic (here with some roundoff error from the summary statistics):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \approx \frac{12.44 - 6.6}{3.638/\sqrt{18}} \approx 6.81$$

- *P*-value, as the area to the right of $t = 6.81$ under the $t(17)$ distribution, because the alternative is one-sided in the greater-than direction. Software commands such as tcdf(6.81,1E99,17) in the TI-83 calculator or = T.DIST.RT(6.81,17) in Excel give a *P*-value of 1.5E−6 (or 0.0000015). Table C tells us only that the one-sided *P*-value must be less than 0.0005 because the test statistic $t = 6.81$ is greater than the largest critical value 3.965 listed for the $t(17)$ distribution.

$\text{df} = 17$

t^*	3.965
One-sided *P*	.0005

These computations are very similar to those we performed in Chapter 14. The difference is that using the sample standard deviation $s = 3.638\ \text{m/s}$ as an approximation for the unknown value of the population standard deviation σ forces us to use the $t(17)$ distribution rather than the standard Normal distribution to compute the test *P*-value.

Any technology will compute the sample mean $\bar{x}$ (estimate), the *t* statistic, and the *P*-value from the data and stated hypotheses. However, outputs vary to some extent in how they report the *P*-value, as Figure 17.6 illustrates. Some technologies, such as R and the TI-83 graphing calculator, give the exact value of *P* even when it is very small. Others round *P*-values to three or four decimal places: Minitab's output "*P* 0.000" actually means that $P<0.0005$,$P < 0.0005$, and both JMP and CrunchIt! report that $P<0.0001$.$P < 0.0001$. Some technologies ask for the direction of the alternative hypothesis (here CrunchIt!, Minitab, R, and the TI-83) and provide the corresponding *P*-value. Others, like JMP here, give a two-sided *P*-value called "Prob > | t |" as well as two one-sided *P*-values called "Prob > t" and "Prob < t." In such cases, you need to select the *P*-value that corresponds to your stated alternative hypothesis (here "Prob > t" because the alternative is one-sided in the greater-than direction).

The step-by-step computations involved in a one-sample *t* test are slightly different from those for a one-sample $z$$z$ test for a population mean. However, all issues related to the interpretation of the *P*-value and its limitations are exactly the same, so that the issues discussed in Chapters 14 and 15 apply equally well to this chapter, as Exercises 17.12 and 17.13 illustrate.

APPLY YOUR KNOWLEDGE

Many exercises from this point onward ask you to give the P-value of a t test. If you have suitable technology, give the exact P-value. Otherwise, use Table C to give an approximate P-value.

17.9 Trout habitat. Do the data of Exercise 17.8 give good reason to think that the springtime water in the tributary water basin around the Shavers Fork watershed is not neutral (a neutral pH is the pH of pure water, $\mathrm{pH}=7$$\mathrm{pH} = 7$)? Follow the four-step process as illustrated in Example 17.3.

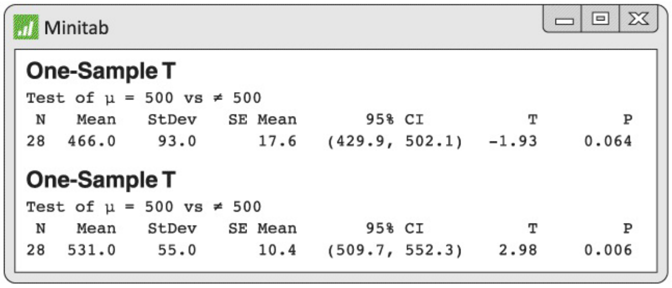
17.10 An attractive body, continued. Exercise 17.6 described a study of the attractiveness of body types. From Barbie dolls to runway models, women in Western countries are exposed to unrealistically thin and arguably unhealthy body standards for their gender. Does the study provide evidence that young Caucasian women in British universities, on average, aim for an unhealthy ideal body type (corresponding to a BMI less than 18.5)?

a. State the corresponding null and alternative hypotheses.
b. Obtain the test statistic and *P*-value then conclude in context.

17.11 Counting steps. A study examined the accuracy of consumer technology used for tracking physical activity. The report states that when a random sample of 28 healthy adults each walked exactly 500 steps on a treadmill, the Fitbit Flex reported an average of 466 steps with a standard deviation of 93 steps. The iPhone 5s Moves app, by comparision, reported an average of 531 steps with a standard deviation of 55 steps.[9] Do these results provide significant evidence that these fitness tools report inaccurate numbers of steps on average? Figure 17.7 gives the Minitab output for each technology.

a. State the null and alternative hypotheses for each technology.
b. Give the test statistic and *P*-value for the Fitbit Flex and conclude in context.
c. Give the test statistic and *P*-value for the iPhone 5s Moves app and conclude in context.

wavebreakmedia/Shutterstock.com



```
One-Sample T
Test of μ = 500 vs ≠ 500
 N    Mean    StDev   SE Mean        95% CI            T        P
28   466.0     93.0      17.6   (429.9, 502.1)   -1.93    0.064

One-Sample T
Test of μ = 500 vs ≠ 500
 N    Mean    StDev   SE Mean        95% CI            T        P
28   531.0     55.0      10.4   (509.7, 552.3)    2.98    0.006
```

Figure 17.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.7 Software output (Minitab) for two wearable fitness devices.

17.12 An attractive body, continued. Exercise 17.6 provided a 95% confidence interval for the mean ideal BMI of young Caucasian women in British universities. In Exercise 17.10 you conducted a hypothesis test for the same data.

a. Explain briefly how the confidence interval could be used to reject or fail to reject your null hypothesis.
b. A confidence interval provides more useful information than a *P*-value when reaching a conclusion in context. What can be concluded about women's mean ideal BMI in this population from the confidence interval but not from the *P*-value?

17.13 Counting steps, continued. Go back to Exercise 17.11 on the accuracy of consumer technology used for tracking physical activity.

a. Figure 17.7 gives a different *P*-value for the Fitbit Flex and for the iPhone 5s Moves app. Considering the computations required to obtain these *P*-values, what was the biggest contributor to having such different *P*-values? Use this example to explain why it is not possible to conclude about effect size simply from a *P*-value.
b. The study described in Exercise 17.11 examined the accuracy of 10 different fitness tools (wearable devices and apps) in the same manner. What would be a major issue with running 10 different *t* tests to examine their accuracy?

MATCHED PAIRS *t* PROCEDURES

Comparative studies are more informative than investigations of a single condition. For that reason, one-sample inference is less common than comparative inference. There is, however, a widely used design to compare two conditions that makes use of one-sample procedures. In a **matched pairs design,** subjects are matched in legitimate (non-arbitrary) pairs, and each treatment is given to one subject in each pair. Another situation calling for matched pairs arises when we record two responses for each subject (repeated measures), as in before-and-after observations or when two treatments are given at different times. We first introduced the matched pairs design in Chapter 7.

MATCHED PAIRS *t* PROCEDURES

To compare the quantitative responses to the two conditions in a matched pairs design, find the difference between the responses within each pair. Then apply the one-sample *t* procedures to these paired differences.

The parameter μμ in a matched pairs *t* procedure is the mean of the pairwise difference in the responses to the two conditions within matched pairs of subjects in the entire population. (The notations μdiffμ_{diff} or μdμ_d are sometimes used to reflect the matched pairs design.)

EXAMPLE 17.4 A treatment for progeria

STATE: As described in Example 17.3, investigators in a clinical study recorded PWV, a measure of vascular stiffness, in 18 children diagnosed with progeria. The objective of the study was to test the effectiveness of treatment with the drug lonafarnib, an enzyme inhibitor that interferes with the mutated protein progerin responsible for the syndrome. The 18 children in Example 17.3 all received a daily dose of lonafarnib for two years. At the end of the two-year treatment, PWV was assessed once more. Table 17.1 gives the PWV values (in m/s) at the beginning (“Untreated”) and at the end (“Treated”) of the study for each child. Is there evidence that a two-year treatment with lonafarnib lowers PWV, on average, in children with progeria?

TABLE 17.1 PWV (m/s) of 18 children with progeria, before and after treatment

Child	Untreated	Treated	Differerence	Child	Untreated	Treated	Differe
1	18.8	12.6	− 6.2	10	12.9	7.5	− 5.4
2	17.6	10.8	− 6.8	11	12.4	9.4	− 3.0
3	17.5	10.1	− 7.4	12	10.1	6.4	− 3.7
4	16.0	10.1	− 5.9	13	9.3	9.0	− 0.3
5	14.8	9.2	− 5.6	14	9.1	6.2	− 2.9
6	14.1	7.6	− 6.5	15	8.3	7.2	− 1.1
7	13.7	10.8	− 2.9	16	7.9	7.3	− 0.6
8	13.1	7.7	− 5.4	17	8.3	5.7	− 2.6
9	12.9	6.8	− 6.1	18	7.2	9.1	1.9

PLAN: Take μμ to be the mean of the variable “difference” (PWV treated minus PWV untreated) in the population of children with progeria. The null hypothesis says that the drug lonafarnib has no effect (paired differences average to zero), and HaH_a says that the drug is effective so that PWV values are smaller after treatment, on average. Therefore we test the hypotheses

H0:μ=0Ha:μ<0

$$\begin{aligned}H_0 &: \mu = 0\\ H_a &: \mu < 0\end{aligned}$$

SOLVE: We are willing to regard the 18 subjects as an SRS from the population of all children with progeria. To analyze the data, subtract the untreated PWV values from the treated PWV values *for each child.* A negative difference shows that a subject’s PWV has decreased after treatment. The 18 differences (listed in the “Difference” column in Table 17.1) form one random sample from the population of paired differences in all children with progeria who could be treated with lonafarnib. Figure 17.8 is a dotplot of the 18 differences. The distribution is very mildly skewed, but with 18 values, the *t* procedures will still be valid.

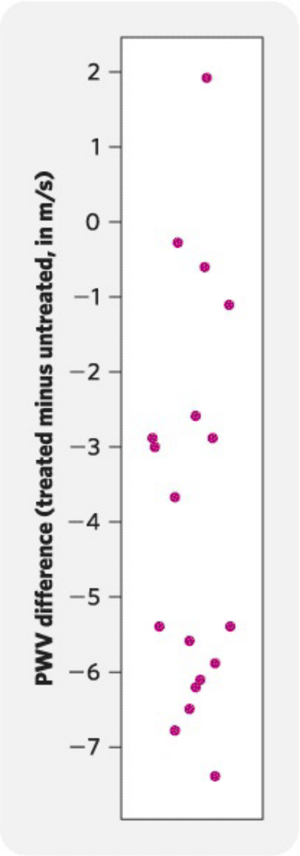


Figure 17.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.8 Dotplot (with jitter) of PWV differences for 18 children with progeria.

Here is the Minitab output for the appropriate *t* test:

```
Paired T for Treated - Untreated
               N     Mean   StDev   SE Mean
Treated       18    8.528   1.885     0.444
Untreated     18   12.444   3.638     0.857
Difference    18   -3.917   2.644     0.623

T-Test of mean difference = 0 (vs < 0): T-Value = -6.29 P-Value = 0.000
```

The test *P*-value is reported as “0.000” because Minitab rounds *P*-values to three decimal places, which indicates that P<0.0005$P < 0.0005$ (highly statistically significant).

CONCLUDE: The data provide very strong evidence (P<0.0005) $(P < 0.0005)$ that a two-year treatment with lonafarnib lowers PWV (improves vascular stiffness) in children with progeria, on average. In

> this study, the 18 children benefited from an average reduction in PWV of 3.9 m/s from their earlier average PWV of 12.4 m/s.

The computations performed by software in Example 17.4 involve the following steps:

- Summary statistics: x-≈-3.917,$\bar{x} \approx -3.917,$ s≈2.644,$s \approx 2.644,$ and n=18 $n = 18$
- Degrees of freedom of the *t* distribution: n-1=18-1=17$n - 1 = 18 - 1 = 17$
- Test statistic (with some roundoff error when using summary statistics):

t=x--0s/n≈-3.917-02.644/18≈-6.29

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} \approx \frac{-3.917 - 0}{2.644/\sqrt{18}} \approx -6.29$$

- *P*-value, as the area left of t=-6.29$t = -6.29$ under the t(17)$t\,(17)$ distribution, because the alternative is one-sided in the less-than direction. Software commands such as tcdf(-1E99,-6.29,17) in the TI-83 calculator or pt(-6.29,17) in R give a *P*-value of 4E−6 (or 0.000004). Table C tells us only that the one-sided *P*-value must be less than 0.0005 because, in absolute terms, t=6.29$t = 6.29$ is greater than the largest critical value 3.965 listed for the t(17)$t\,(17)$ distribution.

df=17$\text{df} = 17$

t*t^*	3.965
One-sided *P*	.0005

Remember that a *P*-value, no matter how small, does not indicate whether an effect is large or important. A confidence interval is better suited to discuss effect size. Here is Minitab’s output when the data of Example 17.4 are used to obtain a 95% confidence interval for the population mean change in PWV:

```
95% CI for mean difference: (-5.231, -2.602)
```

Thus, we are 95% confident that the true mean reduction in PWV after taking lonafarnib for two years is somewhere between 2.6 and 5.2 m/s, in the population of children diagnosed with progeria. While the confidence interval tells us about the expected size of the effect in the target population, it doesn't tell us whether this kind of effect is clinically meaningful. Only a person knowledgeable about PWV and the physiology of children diagnosed with progeria can determine the clinical relevance of a mean reduction between 2.6 and 5.2 m/s. In addition, any benefit from a PWV reduction would need to be weighed against any potential side effects of the treatment. Those are very important considerations, but they are not addressed by statistical inference.

Example 17.4 illustrates how to turn matched pairs data into single-sample data by calculating a difference within each pair.[10] We are making inferences about a single population—that is, the population of all differences within matched pairs. *It is incorrect to ignore the matching and analyze the data as if we had two samples:* children treated with lonafarnib and children left untreated. Inference procedures for comparing two samples assume that the samples are selected independently of each other, which is not true when the same subjects are measured twice. The proper analysis depends on the design used to produce the data. We will describe the procedures used for two independent samples in the next chapter.

APPLY YOUR KNOWLEDGE

17.14 Water diet. Water consumption before a meal has been shown to help reduce the caloric intake from that meal, but does it translate into actual weight loss? An experiment randomly assigned overweight and obese individuals aged 55 to 75 years to one of two diet groups for 12 weeks: a hypocaloric diet supplemented with 16 ounces of bottled water prior to each of the three daily meals (water group) or a hypocaloric diet alone (nonwater group).[11]

a. Each subject had his or her body mass index (BMI) computed before and after the 12-week diet. For the 23 subjects assigned to the water group, the mean BMIs before and after treatment were 32.16 and 29.5, respectively. Explain briefly why a matched pairs procedure must be used to assess whether the combination of a hypocaloric diet and pre-meal water intake significantly lowers BMI.

b. The 25 patients assigned to the nonwater group had mean BMIs before and after the 12-week diet of 31.8 and 29.9, respectively. Explain briefly why we cannot use a matched pairs procedure to test the hypothesis that the combination of a hypocaloric diet and pre-meal water intake is significantly more effective than a hypocaloric diet alone in lowering BMI.

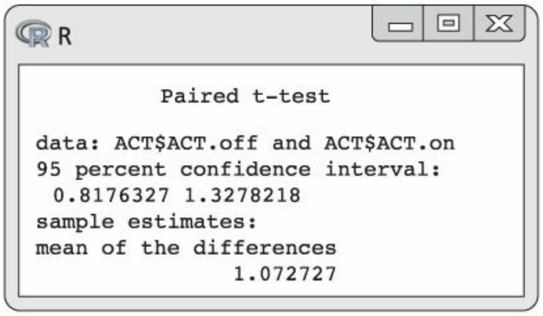
17.15 Essential tremor. Essential tremor is a neurological movement disorder characterized by involuntary rhythmic movement that typically interferes with the full use of the arms and hands. Example 7.8 (page 175) described a pilot experiment examining the effectiveness of a noninvasive handheld device using active cancellation of tremor (ACT) technology to stabilize tremor-induced motion in patients diagnosed with essential tremor. Here is the tremor amplitude measured (in centimeters) with a precision accelerometer for each of 11 subjects when performing a spoon-use task with the ACT device turned, in random order, once on and once off.

Subject	1	2	3	4	5	6	7	8	9	10	11
ACT off	1.2	1.6	2	1	1.7	1.9	0.8	1.7	1.6	0.5	1.4
ACT on	0.3	0.2	0.6	0.2	0.5	0.4	0.2	0.6	0.3	0.2	0.1

a. Figure 17.9 shows the R output for the 95% confidence interval for the mean difference in tremor amplitude based on these data. Interpret this confidence interval in context.

b. Compute the 11 paired differences in tremor amplitude (off minus on), then obtain their mean and standard deviation.

c. The critical value for a 95% confidence level under the t(10)$t\,(10)$ distribution is 2.228. You may want to verify that this value is correct from either Table C in the back of the book or statistical software. Using the formula on page 422, compute the margin of error of a 95% confidence interval for the mean difference in tremor amplitude. Verify that the resulting confidence interval agrees with the R output.



```
          Paired t-test

data: ACT$ACT.off and ACT$ACT.on
95 percent confidence interval:
 0.8176327 1.3278218
sample estimates:
mean of the differences
               1.072727
```

Figure 17.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 17.9 Software output (R) for the essential tremor matched pairs experiment.

17.16 A bionic pancreas. A study tested the effectiveness of a bionic pancreas for automated glucose management among patients with type 1 diabetes. In random order, each patient received therapy with a bionic pancreas for 5 days and therapy with an insulin pump for 5 days. Table 17.2 lists the plasma glucose level (in milligram per deciliter, mg/dl) attained in each condition for each of 32 adolescent patients.[12] Is there evidence that, on average, a bionic pancreas helps patients attain a lower plasma glucose level than an insulin pump? Follow the four-step process.

TABLE 17.2 Glucose level (mg/dl) of 32 adolescents with type 1 diabetes conditions, administered to each subject in random order

Subject	Control	Bionic	Subject	Control	Bionic	Subject	Control	Bion
1	220	136	9	159	154	17	222	138
2	202	142	10	95	136	18	181	137
3	188	149	11	119	152	19	169	142
4	189	147	12	122	117	20	163	125
5	179	140	13	133	151	21	156	144
6	172	134	14	138	119	22	153	136
7	164	132	15	144	156	23	152	156
8	161	139	16	157	151	24	129	136

ROBUSTNESS OF *t* PROCEDURES

We began this chapter by considering the generic conditions for inference. We now examine these conditions in the context of actual data. The *t* confidence interval and test are exactly correct when the distribution of the population is exactly Normal. Of course, no real data are exactly Normal. Therefore, the usefulness of the *t* procedures in practice depends on how strongly they are affected by lack of Normality.

ROBUST PROCEDURES

A confidence interval or hypothesis test is called **robust** if the confidence level or *P*-value does not change very much when the conditions for use of the procedure are violated.

Fortunately, the *t* procedures are quite robust against non-Normality of the population except when outliers or strong skewness are present. (Skewness is more serious than other kinds of non-Normality in symmetric distributions.) As the sample size n increases, the central limit theorem ensures that the distribution of the sample mean $\bar{x}$ becomes more nearly Normal and that the *t* distribution becomes more accurate for critical values and *P*-values of the *t* procedures.

The presence of one or more outliers in the sample data is often more problematic. Outliers could reflect a strongly skewed population distribution or they could simply be unusual observations. They may even be data entry or experimental mistakes, as discussed in Chapter 2 (page 55). Another issue is that $\bar{x}$ and s are not resistant to outliers, so that one extreme value could have a noticeable impact on the computations needed to obtain a *t* confidence interval or *P*-value. Therefore, *the t procedures are not robust against outliers unless the sample is sufficiently large.*

CAUTION

Always plot the data to check for skewness and outliers before you use the *t* procedures, especially for smaller samples. For most purposes, you can safely use the one-sample *t* procedures when $n \geq 15$ unless an outlier or quite strong skewness is present. Here are practical guidelines for inference about a single mean.[13]

USING THE *t* PROCEDURES

- The condition that the data can be considered an SRS from the population of interest is more important than the condition that the population distribution is Normal.
- *Small sample size less than 15:* Use *t* procedures if the data appear close to Normal (at least roughly symmetric, single peak, no outliers). If the data are skewed or if outliers are present, do not use *t*.
- *Moderate sample size at least 15:* The *t* procedures can be used except in the presence of outliers or strong skewness.
- *Large sample size:* The *t* procedures can be used even for strongly skewed distributions when the sample is large, roughly $n \geq 40$.$n \geq 40$.

EXAMPLE 17.5 Can we use *t*?

Figure 17.10 shows plots of several data sets, all of which can be considered random samples from their respective populations. For which of these can we safely use the *t* procedures?

- Figure 17.10(a) is a dotplot of the sizes (in cubic centimeters, cm^3) of 11 acorns from oak trees on the U.S. Pacific coast.[14] This sample contains a clear and extreme outlier at roughly 17 cm 3.^{3}. With an outlier in a sample this small $(n=11)$,$(n = 11)$, we cannot trust the *t* procedures.
- Figure 17.10(b) is a dotplot of the lengths (in cm) of 15 needles from Aleppo pine trees in California.[15] The data are mildly skewed to the right and there are no outliers. With such a mild deviation from Normality and a moderate sample size $(n=15)$,$(n = 15)$, we can safely use the *t* procedures.
- Figure 17.10(c) is a histogram of the weights (in grams, g) of roughly

500 brown treesnakes on the island of Guam, where this invasive species is causing huge environmental and financial damage.[16] The data are clearly skewed to the right and there is one extreme outlier at a weight of a little more than 650 g. With such a large sample size (n≈500),$(n \approx 500)$, the skew is not a problem for the *t* procedures. The issue with the outlier is that it has more influence than any other value in the data set. For instance, removing the lightest snake in the data set (23 g) changes x-$\bar{x}$ by only 0.2 g, but removing the heaviest snake (663 g) changes x-$\bar{x}$ by 1.1 g (not a large change, but more substantial than 0.2 g). We can safely use the *t* procedures here but further thought should be given to the outlier, as outlined in the Chapter 2 discussion (page 55).

- Figure 17.10(d) is a histogram of the widths (in millimeters, mm) of roughly 800 cocoons from wild bees in Sweden.[17] This distribution is obviously bimodal, reflecting the fact that females (who may eventually become queens) typically have larger cocoons than male workers. The central limit theorem tells us that, for a population distribution of *any* shape, the sampling distribution of x-$\bar{x}$ will be approximately Normal with a large enough sample size n.n. Therefore, with such a large sample (n≈800),$(n \approx 800)$, the bimodal aspect of this data set is not a problem in and of itself for the *t* computations. However, it would be misleading to provide just one measure of center for a distribution that is clearly bimodal. The “center” of any bimodal distribution (lumping two subsets) would be influenced both by the center of each subset and by the relative abundance of each subset, rendering the summary essentially meaningless. Even though the sample sizeis very large, the *t* procedures are not appropriate here because the data are clearly bimodal.

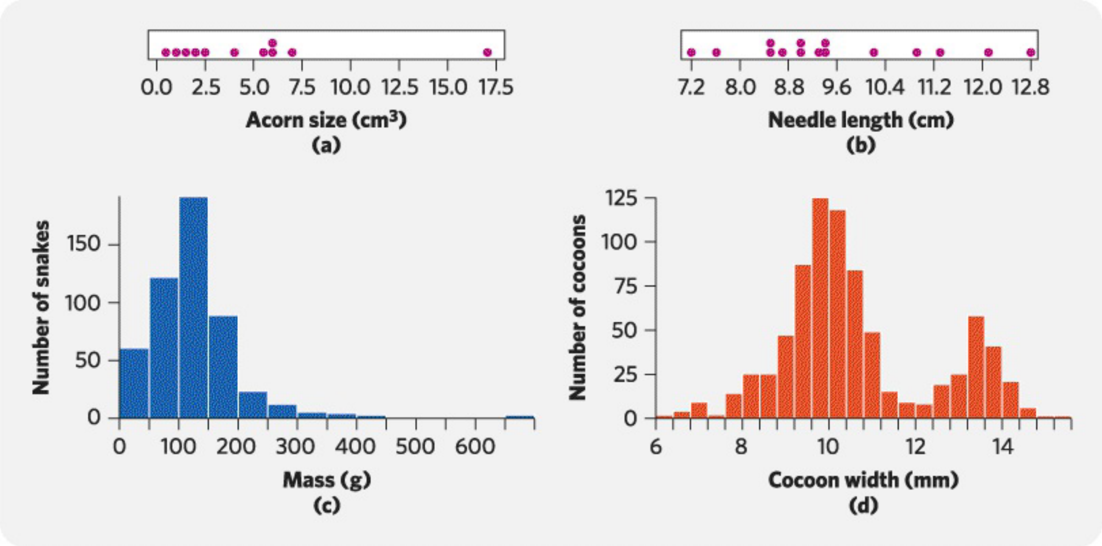


Figure 17.10

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 17.10 Can we use *t* procedures for these data? (a) Volumes of 11 oak tree acorns. No, there is a clear outlier and the sample size is small. (b) Lengths of 15 Aleppo pine needles. Yes, the sample is large enough to overcome the mild skewness. (c) Weights of roughly 500 brown treesnakes. Yes, the very large sample size is enough to overcome the pronounced skewness and the extreme outlier, although the impact of the outlier should be given careful consideration. (d) Widths of roughly 800 bee cocoons. No, because the distribution is clearly bimodal, so it cannot be reasonably described by just one mean.

Narelle Power/Alamy Stock Photo

APPLY YOUR KNOWLEDGE

17.17 Mercury levels in pregnant women. Exercise 2.4 described a study of blood mercury levels (in micrograms per liter, μg/l) recorded in 4134 pregnant British women enrolled in the Avon Longitudinal Study of Parents and Children. This cohort aimed to enroll over a two-year period all pregnant women in a geographical area in the United Kingdom, including both rural and urban parts of the area. The published findings include the following statement: “Blood mercury levels ranged from 0.17 to 12.8 μg/l. The 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles were 0.81, 0.99, 1.35, 1.86, 2.52, 3.33, and 4.02 μg/l, respectively.”

a. Can this data set be reasonably considered as a representative sample of a larger population? Briefly argue why or why not.
b. The numerical summary provided suggests that the distribution of blood mercury levels in this sample is skewed to the right, possibly with high outliers. Would the lack of Normality in this data set affect the validity of a *t* procedure? Briefly explain your answer.
c. Which of the two issues addressed in parts a and b would be of greater concern if we wanted to use a *t* procedure for these data?

17.18 Sitting versus squatting. Squatting for defecation continues to be the traditional position in most of Asia and Africa. In contrast, the practice of sitting on toilet seats has proliferated in Western countries since the 19th century with the development of the sewage system. This difference, along with differences in diet, might explain the higher rate of hemorrhoids and constipation in Western countries. A recent study asked if body position affects defecation in humans. Researchers recruited 6 healthy adult volunteers and took X-rays of their intestinal systems to measure the anorectal angle (in degrees). Note that, as with a water hose, an angle of 180 degrees should provide the least amount of flow disruption. For each subject, an X-ray was taken once in a sitting position and once in a squatting position. Here are the findings:[18]

Subject	Sitting	Squatting
1	109	122
2	73	127
3	117	141
4	77	103
5	98	121
6	125	142

a. State the null and alternative hypotheses. Explain why the investigators used a two-sided alternative.
b. Carry out a test and report your conclusion in simple language.
c. The investigators used the test you just carried out. Any use of the *t* procedures with samples of this size is risky. Why?

Oote Boe 1/Alamy Stock Photo

CHAPTER 17 SUMMARY

- Tests and confidence intervals for the mean μμ of a Normal population are based on the sample mean x-$\bar{x}$ of an SRS. Because of the central limit theorem, the resulting procedures are approximately correct for other population distributions when the sample is large enough.

- The standardized sample mean is the **one-sample** zz **statistic**:

  z=x--μσ/n

  $$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- In practice, we rarely know σ.$\sigma.$ We can, however, replace the standard deviation σ/n$\sigma/\sqrt{n}$ of x-$\bar{x}$ with the **standard error** s/n$s/\sqrt{n}$ and get the **one-sample *t* statistic**:

  t=x--μs/n

  $$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

  This *t* statistic has the *t* **distribution** with n-1$n-1$ degrees of freedom.

- There is a *t* distribution for every positive **degrees of freedom**. All are symmetric distributions similar in shape to the standard Normal distribution. The *t* distribution approaches the N(0,1)$N(0, 1)$ distribution as the degrees of freedom increases.

- **Technology** (statistical software, graphing calculators, applets) can be used to automate the computations, based on the t(n-1)$t(n-1)$ distribution, required to obtain a confidence interval for the mean μμ of a population or to test a set of null and alternative hypotheses about μ.$\mu.$

- A level *C* **confidence interval for the mean** μμ of a Normal population is:

  x-±t*sn

  $$\bar{x} \pm t^{*} \frac{s}{\sqrt{n}}$$

  The **critical value** t*t^{*} is the value such that the t(n-1)$t(n-1)$ density curve has area *C* between -t*$-t^*$ and t*.t^*.

- **A hypothesis test** for H0:μ=μ0$H_0 : \mu = \mu_0$ is based on the value of the *t* statistic when H0H_0 is true. Use the test *P*-value, computed as an area under the t(n-1)$t\,(n-1)$ distribution, to decide whether to reject or fail to reject H0.H_0.
- Use these one-sample procedures to analyze **matched pairs** data by first taking the difference within each matched pair to produce a single sample.
- The *t* procedures are quite **robust** when the population is non-Normal, especially for larger sample sizes. The *t* procedures are useful for non-Normal data when n≥15$n \geq 15$ unless the data show outliers or strong skewness (in which case, a larger nn might be needed).

THIS CHAPTER IN CONTEXT

Chapters 14 and 15 introduced the idea of statistical inference for a population mean. We relied on the assumption that we knew the value of the population standard deviation σ.σ. Knowing the value of σσ makes the computations for a confidence interval or a *P*-value easier because we can use a Normal distribution for the sampling distribution of the sample mean (detailed in Chapter 13). In this chapter, we describe how the *t* distribution can be used instead when σσ is unknown—a much more common situation by far. The reasoning behind statistical inference about the population mean is similar; only the details of the computations differ.

Optional sections in Chapter 15 showed how to determine the sample size for a desired margin of error or for a desired test power. This can be done for *t* procedures as well, although the actual computations are more complicated because the degrees of freedom for the *t* procedures also depends on the sample size (df =n-1).$= n - 1$). In practice, these computations are left to dedicated statistical software.

This chapter also explores the concept of inference for a matched pairs design (introduced in Chapter 7), when two conditions are compared within pairs—either by using meaningful pairs of individuals (for example, siblings) or by obtaining a pair of measurements from each individual in the sample. In Chapter 18 we will describe a *t* inference procedure that lets us compare the means of two populations when the data have been collected using two independent samples. In Chapters 24 and 26 we will introduce a new method (ANOVA) that allows us to compare the means of more than two populations.

Before using a *t* procedure or ANOVA on quantitative data, you must check whether the procedure's assumptions are met. This requires plotting your data and identifying possible skew or outliers, as you first learned in Chapters 1 and 2. Optional Chapter 27 on nonparametric tests describes alternative inference procedures that can be used when the assumptions for a *t* procedure or ANOVA are not met—most commonly when sample sizes are small.

CHECK YOUR SKILLS

17.19 We prefer the *t* procedures to the z procedures for inference about a population mean because

a. z can be used only for large samples.
b. z requires that you know the population standard deviation σ.
c. z requires that you can regard your data as an SRS from the population.

Use the following information for Exercises 17.20 through 17.22. Data on the blood cholesterol levels (in milligrams per deciliter) of a random sample of 24 lab rats give $\bar{x} = 85$ *and* $s = 12$.

17.20 A 95% confidence interval for the mean blood cholesterol of lab rats is

a. 78.7 to 91.3.
b. 79.9 to 90.1.
c. 80.2 to 89.8.

17.21 The correct interpretation for this confidence interval is that we are 95% confident that

a. the 24 rats in this sample have blood cholesterol levels within this interval of values.
b. 95% of all rats have blood cholesterol levels within this interval of values.
c. the mean blood cholesterol level of all rats lies within this interval of values.

17.22 Which of the following would be most worrisome for the validity of this confidence interval?

a. The presence of a clear outlier in the raw data
b. A mild skew in the dotplot of the raw data
c. Not knowing the population standard deviation σ

Use the following information for Exercises 17.23 through 17.25.

Researchers recruited 16 young adults to participate in a study of “Wii sport” energy expenditure. Participants got a chance to learn the activity and then performed it for 10 minutes while hooked up to monitors measuring the calories they burned (in kilocalories per minute, Cal/min).[19] *For Wii tennis, participants burned on average 5.3 Cal/min with a standard deviation of 1.8 Cal/min. Is this evidence that playing Wii tennis burns less calories on average than playing actual tennis, which is generally thought to burn on average 8.1 Cal/min?*

17.23 Which of the following is the appropriate null hypothesis?

a. H0: μWii tennis=0$H_0 : \mu_{Wii\ tennis} = 0$

b. H0: μWii tennis=5.3$H_0 : \mu_{Wii\ tennis} = 5.3$

c. H0: μWii tennis=8.1$H_0 : \mu_{Wii\ tennis} = 8.1$

17.24 What is the value of the *t* statistic?

a. -6.22-6.22

b. -2.80-2.80

c. -1.56-1.56

17.25 The *P*-value for this test is

a. greater than 0.05.

b. between 0.05 and 0.001.

c. less than 0.001.

Use the following information for Exercises 17.26 and 17.27. Does reading on a light-emitting e-reader at bedtime increase how long it takes to fall asleep? A sample of 12 healthy adults slept in the lab on two different nights. In random order, participants read for 30 minutes in a print book one night and with a light-emitting e-reader the other night. Scalp electrodes were used to measure how long (in minutes) it took participants to reach a deep sleep stage. On average, the subjects took 12.3 minutes longer to reach deep sleep with the e-reader than with the print book (P=0.009).$(P = 0.009)$.[20]

17.26 Which of the following is the appropriate null hypothesis here?

a. H0: μe-reader=12.3$H_0 : \mu_{e-reader} = 12.3$

b. H0: μdifference=0$H_0 : \mu_{difference} = 0$

c. H0: μe-reader-μbook=0$H_0: \mu_{e-reader} - \mu_{book} = 0$

17.27 What can we conclude for a population of individuals similar to those studied?

a. There is strong evidence (P=0.009)$(P = 0.009)$ that, on average, it takes longer to fall deeply asleep at night after reading on a light-emitting e-reader than after reading from a book.
b. There is strong evidence (P=0.009)$(P = 0.009)$ that, on average, it takes 12.3 minutes longer to fall deeply asleep at night after reading on a light-emitting e-reader than after reading from a book.
c. It takes much longer, on average (P=0.009),$(P = 0.009)$, to fall deeply asleep at night after reading on a light-emitting e-reader than after reading from a book.

17.28 Because the *t* procedures are robust, the most important condition for their safe use is that

a. the population standard deviation σσ is known.
b. the population distribution is exactly Normal.
c. the data can be regarded as an SRS from the population.

CHAPTER 17 EXERCISES

17.29 Sharks. Great white sharks are big and hungry. Here are the lengths in feet of 44 great whites:[21]

18.7	12.3	18.6	16.4	15.7	18.3	14.6	15.8	14.9	17.6	12.1
16.4	16.7	17.8	16.2	12.6	17.8	13.8	12.2	15.2	14.7	12.4
13.2	15.8	14.3	16.6	9.4	18.2	13.2	13.6	15.3	16.1	13.5
19.1	16.2	22.8	16.8	13.6	13.2	15.7	19.7	18.7	13.2	16.8

a. Examine these data for shape, center, spread, and outliers. The distribution is reasonably Normal except for one outlier in each direction. Because these outliers are not extreme and preserve the symmetry of the distribution, use of the *t* procedures is safe with 44 observations.

b. Give a 95% confidence interval for the mean length of great white sharks. Based on this interval, is there significant evidence at the 5% level to reject the claim "Great white sharks average 20 feet in length"?

c. Before accepting the conclusions of part b, you need more information about the data. What would you like to know?

17.30 Alcohol in wine. The alcohol content of wine depends on the grape variety, the way in which the wine is produced from the grapes, the weather, and other influences. Here are data on the percent of alcohol in wine produced from the same grape variety in the same year by 48 winemakers in the same region of Italy:[22]

12.86	12.88	12.81	12.70	12.51	12.60	12.25	12.53	13.49	12.84
12.93	13.36	13.52	13.62	12.25	13.16	13.88	12.87	13.32	13.08
13.50	12.79	13.11	13.23	12.58	13.17	13.84	12.45	14.34	13.48
12.36	13.69	12.85	12.96	13.78	13.73	13.45	12.82	13.58	13.40
12.20	12.77	14.16	13.71	13.40	13.27	13.17	14.13		

a. Make a histogram of the data, using class width 0.25. The shape of the distribution is a bit irregular, but there are no outliers or strong skewness. There is no reason to avoid use of *t* procedures for $n=48$.

$n = 48.$

b. Give a 95% confidence interval for the mean alcohol content of wine of this type.
c. Based on your confidence interval, is the mean alcohol content significantly different at the $\alpha=0.05$$\alpha = 0.05$ level from 12 percent? From 13 percent?

17.31 Wood lice. Students in a physiology lab collected 50 wood lice (“pill bugs”) from the university grounds. Regard these specimens as a random sample of wood lice on the university grounds. Here are their lengths in millimeters:[23]

11.0	11.1	10.1	10.8	9.2	9.0	11.0	9.5	8.0	14.6
9.1	9.7	12.0	12.5	13.1	11.7	8.1	8.7	11.9	10.3
8.1	8.3	8.2	12.1	12.3	10.9	8.9	7.5	9.0	8.2
8.3	9.0	7.9	10.5	10.0	9.8	8.9	11.0	11.0	10.9
11.0	11.5	11.0	13.9	9.3	8.5	10.2	8.3	9.1	7.0

Verify that there are no outliers in the data. What is a 95% confidence interval for the mean body length in this population?

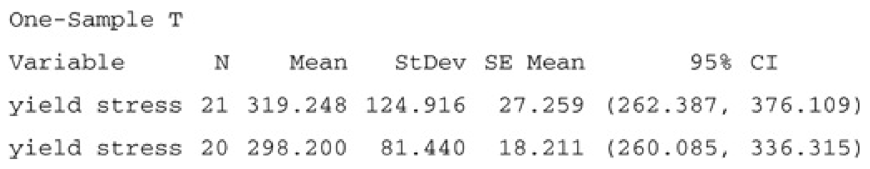
17.32 More on spider silk. Spider silk is the strongest known material, natural or human-made, on a weight basis. A study examined the mechanical properties of spider silk using a sample of 21 female golden orb weavers, *Nephila clavipes*. It is reasonable to regard these spiders as a random sample of their population. Here are data (in megapascals, MPa) on silk yield stress, which represents the amount of force per unit area needed to achieve permanent deformation of the silk strand:[24]

164.0	478.7	251.3	351.7	173.0	448.9	300.6
362.0	272.4	740.2	329.0	327.2	270.5	332.1
288.8	176.1	282.2	236.1	358.2	270.5	290.7

a. Plot the data in a dotplot and describe the distribution of yield stress values in this sample.
b. We would like to obtain a 95% confidence interval for the mean yield

stress of golden orb weaver silk. Are the conditions for a *t* confidence interval met? Explain briefly.

c. Here is the Minitab output for the data with and without the high outlier:



```
One-Sample T

Variable       N     Mean    StDev  SE Mean        95% CI

yield stress  21  319.248  124.916   27.259  (262.387, 376.109)

yield stress  20  298.200   81.440   18.211  (260.085, 336.315)
```

Does the outlier have a noticeable influence on the confidence interval? How to handle outliers is a complex challenge that requires detailed investigation. Refer to the discussion of outliers in Chapter 2 (page 55) for more information.

17.33 Conditions for inference. Exercise 17.1 gave the summary data (x-=27.8,s=32.2 spikes per second)$(\bar{x} = 27.8, s = 32.2 \text{ spikes per second})$ for the baseline activity of a neuron from 72 separate recordings. Neural activity is expressed in number of action potentials per second and cannot be less than 0. You notice that the standard deviation is larger than the mean. What does this suggest about the shape of the distribution? Is it nonetheless acceptable to calculate a *t* confidence interval for the population mean? Explain why or why not.

17.34 Vision screening. Vision problems in young children often go undetected, delaying proper care. Researchers examined a new computer-based eye exam screening test. In the study sample of 175 young children (ages 3 to 7 years), screening times ranged from 23 to 357 seconds (sec), with a mean of 84 sec and a standard deviation of 43 sec.[25] The software is freely available and could be useful for screening large numbers of elementary and preschool children.

a. Estimate with 95% confidence the mean amount of time it takes to screen a young child with this computer-based eye exam.
b. Check the conditions for the inference procedure you used in part a.

17.35 Hand size. A clever way to determine hand size in three dimensions is to measure the volume (in millimeters, ml) of water displaced when the hand is dipped in a water container. A study used this method to gather the hand volumes of 12 male college students. Here are the measurements:[26]

400 360 420 520 460 350 500 420 450 430 395 400

a. We can consider this an SRS of all male college students. Make a stemplot or a dotplot. Is there any sign of major deviation from Normality?
b. Obtain a 95% confidence interval for the mean hand size and interpret it in context.

17.36 Essential tremor, continued. Exercise 17.15 described a study of the effectiveness of the ACT device for patients diagnosed with essential tremor. Estimate with 95% confidence the mean tremor amplitude in the population of similar patients when using the ACT device. Follow the four-step process.

4 STEP

17.37 Genetic engineering for cancer treatment. Here's a new idea for treating advanced melanoma, the most serious kind of skin cancer: Genetically engineer white blood cells to better recognize and destroy cancer cells, and then infuse these cells into patients. The subjects in a small initial study of this treatment were 11 patients whose melanoma had not responded to existing treatments. One question was how rapidly the new cells would multiply after infusion, as measured by the doubling time (in days). Here are the doubling times:[27]

1.4 1.0 1.3 1.0 1.3 2.0 0.6 0.8 0.7 0.9 1.9

a. Examine the data. Is it reasonable to use the *t* procedures?
b. Give a 90% confidence interval for the mean doubling time. Are you willing to use this interval to make an inference about the mean doubling time in a population of similar patients?

17.38 Genetic engineering for cancer treatment, continued. Another outcome in the cancer experiment described in Exercise 17.37 is measured by looking for the presence of cells that trigger an immune response in the body and so may help fight cancer. The following table gives the counts of active cells per 100,000 cells before and after infusion for 11 subjects. The difference (after minus before) is the response variable.

Before	14	0	1	0	0	0	0	20	1	6	0
After	41	7	1	215	20	700	13	530	35	92	108
Difference	27	7	0	215	20	700	13	510	34	86	108

a. Examine the data. Is it reasonable to use the *t* procedures?

b. If your conclusion in part a is "yes," do the data provide convincing evidence that the count of active cells is higher after treatment?

17.39 DNA on the ocean floor. We typically think of DNA as the stuff that stores the genetic code. It turns out that, outside living cells, DNA mainly occurs on the ocean floor, where it plays an important role in nourishing seafloor life. A random sample of ocean floor specimens from 116 locations around the world gives mean DNA concentration x-=0.2781 $\bar{x} = 0.2781$ and standard deviation s=0.1803$s = 0.1803$ (in grams per square meter).[28]

a. Compute a 95% confidence interval for mean DNA concentration on the ocean floor all around the world.

b. The sample distribution of 116 concentrations is strongly right-skewed. Is it appropriate to compute a *t* interval for these data? Briefly explain your answer.

17.40 Oligofructose and calcium absorption. Nondigestible oligosaccharides are known to stimulate calcium absorption in rats. A double-blind, randomized experiment investigated whether the consumption of oligofructose similarly stimulates calcium absorption in healthy male adolescents 14 to 16 years old. The subjects took a pill for nine days and had their calcium absorption tested on the last day. The experiment was repeated three weeks later. Some subjects received the oligofructose pill in the first round and then a pill containing sucrose (which served as a control)in the second round. The order was switched for the remaining subjects. Here are the fractional calcium absorption data (in percent of intake) for 11 subjects:[29]

Subject	1	2	3	4	5	6	7	8	9	10	11
Control	78.4	76.6	57.4	51.5	49.0	46.6	44.2	42.9	37.2	34.1	24.6
Oligofructose	62.0	95.1	46.5	49.4	89.7	43.8	50.3	51.6	66.6	52.7	54.0

a. Examine the data. Is it reasonable to use the *t* procedures?

b. Do the data support the hypothesis that oligofructose facilitates calcium absorption?

17.41 The placebo effect. The placebo effect is particularly strong in patients with Parkinson's disease. To understand the workings of the placebo effect, scientists measure activity at a key point in the brain when patients receive a placebo that they think is an active drug and also when no treatment is given.[30] The same 6 patients are measured both with and without the placebo, at different times.

a. Explain why the proper procedure to compare the mean response to the placebo with the mean response to the control (no treatment) is a matched pairs *t* test.

b. The six differences (treatment minus control) had x-=-0.326$\bar{x} = -0.326$ and s=0.181.$s = 0.181.$ Is there significant evidence of a difference between treatment and control?

17.42 Radioactive contamination of tuna. Radioactive cesium (cesium-134 and cesium-137) is a waste product of nuclear reactors. A study examined the tissue concentration of radioactive cesium in a random sample of 15 Pacific bluefin tuna, *Thunnus orientalis*, captured off the coast of California four months after the Fukushima (Japan) nuclear reactor meltdown of 2011. Here are the findings, in becquerels per kilogram of dry tissue (Bq/kg):[31]

4.6	6.9	7.9	8.0	8.0	9.2	10.4	10.6
11.1	11.1	11.3	12.3	12.7	14.2	15.6	

Historically, radioactive cesium tissue concentrations in bluefin tuna have been below 2 Bq/kg, on average. Do a complete analysis of the data, including a 95% confidence interval for the mean tissue concentration of radioactive cesium in bluefin tuna off the coast of California four months after the 2011 nuclear meltdown. Use your confidence interval to compare these radioactive cesium levels with the historical average. Follow the four-step process.

17.43 Do chimpanzees collaborate? Humans often collaborate to solve problems. Would a chimpanzee recruit another chimp when solving a problem that requires collaboration? Researchers presented 8 chimpanzee subjects with food outside their cage that they could bring within reach by pulling two ropes, one attached to each end of the food tray. If a chimp pulled only one rope, the rope came loose and the food was lost. Another chimp was available as a partner, but only if the subject unlocked a door joining two cages. (Chimpanzees learn these things quickly.) The same 8 chimpanzee subjects faced this problem in two versions: The two ropes were close enough together that one chimp could pull both (no collaboration needed), or the two ropes were too far apart for one chimp to pull both (collaboration needed). Table 17.3 shows how often in 24 trials each subject opened the door to recruit another chimp as partner.[32] Do the data provide evidence that chimpanzees recruit partners more often when a problem requires collaboration? Follow the four-step process.

TABLE 17.3 Trials (out of 24) in which chimpanzees recruited a partner

Chimpanzee	Collaboration Needed		Difference
	Yes	No	
Namuiska	16	0	16
Kalema	16	1	15
Okech	23	5	18
Baluku	19	3	16
Umugenzi	15	4	11
Indi	20	9	11
Bili	24	16	8
Asega	24	20	4

17.44 Comparing two drugs. Makers of generic drugs must show that they do not differ significantly from the “reference” drugs that they imitate. One aspect in which drugs might differ is the extent of their absorption in the blood. Table 17.4 gives data taken from 20 healthy nonsmoking male subjects for one pair of drugs.[33] This is a matched pairs design. Numbers 1 to 20 were assigned at random to the subjects. Subjects 1 to 10 received the generic drug first, and Subjects 11 to 20 received the reference drug first. In all cases, a washout period separated the two drugs so that the first drug had disappeared from the blood before the subject took the second drug. Do the drugs differ significantly in absorption?

TABLE 17.4 **Absorption extent for two versions of a drug**

Subject	Reference	Generic	Subject	Reference	Generic	Subject	Reference	G
15	4108	1755	20	1709	1878	5	938	
3	2526	1138	17	1829	1682	7	1339	
9	2779	1613	2	2594	2613	14	1262	
13	3852	2254	4	2344	2738	11	1438	
12	1833	1310	16	1864	2302	1	1735	
8	2463	2120	6	1022	1284	19	1020	
18	2059	1851	10	2256	3052			

17.45 Evolution and relative fitness. Can bacteria evolve a preference for the pH of their environment? An evolutionary biologist examined the relative fitness of *Escherichia coli* bacteria grown for 2000 generations (about 300 days) at stressful acidic pH 5.5 and their parental generation grown and preserved at pH 7.2. Both types were later grown together in an acidic medium, and their relative fitness was computed. The experiment was replicated with 6 different lines of *E. coli*, giving the following relative fitness values:[34]

1.24	1.22	1.23	1.24	1.18	1.09

A relative fitness of 1 indicates that both bacteria types are equally fit. A relative fitness larger than 1 indicates that the acid-evolved line is more fit than the parental line kept at neutral pH when both are grown in acidic conditions (that is, the acid-evolved bacteria grew the most). Do the data provide evidence that bacteria evolved in acidic pH are better adapted to acidic conditions? Do a complete analysis, following the four-step process illustrated in Example 17.3.

Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Note that short answers are not available for these exercises to enable their use for class work or assignments.*

17.46 Estimating body fat in men. The data file *Large.Bodyfat* contains data on the percent of body fat, age, weight, height, and 10 body circumference measurements for 252 adult men. In Exercise 8.16 (page 208), you examined the distribution of weights and heights and identified a high outlier in each distribution.

a. The outlier in the distribution of men's heights was a typo. Obtain a 95% confidence interval for the mean height (in inches) of adult men, making sure to exclude the incorrect value.
b. Estimate the mean body weight (in pounds) of adult men with 95% confidence, first using all 252 data points and then without the high outlier. Compare the two intervals. What is the impact of the outlier?

17.47 Elderly health. Exercise 2.42 (page 65) describes a study of calcium or inorganic phosphorus blood levels (both measured in millimoles per liter, mmol/l) in elderly patients. The data file *Large. Calcium* contains the data. Find a 90% confidence interval for the mean calcium level of all elderly men (in the data file, *sex* = 1 corresponds to men); also compute a 90% confidence interval for the mean calcium level of all elderly women (*sex* = 2). Do the same for the inorganic phosphorus levels. Write a short paragraph describing your findings. (Describe each result separately. We will see how to determine if two sample means are significantly different in Chapter 18.)

CHAPTER 18 Comparing Two Means

GUDKOVANDREY/Depositphotos

IN THIS CHAPTER WE COVER...

- Comparing two population means
- Two-sample *t* procedures
- Robustness again
- Avoid the pooled two-sample *t* procedures
- Avoid inference about standard deviations

Using two distinct samples of unrelated individuals to compare two populations or two treatments is one of the most common situations encountered in statistical practice. We call such situations *two-sample problems*.

TWO-SAMPLE PROBLEMS

- The goal of inference is to compare the population responses to two treatments or to compare the characteristics of two populations.
- We have a separate sample for each treatment or each population.

COMPARING TWO POPULATION MEANS

A two-sample problem can arise from a randomized comparative experiment that randomly divides subjects into two groups and exposes each group to a different treatment. Comparing random samples selected separately from two populations is also a two-sample problem. So is comparing two subsets of individuals within a general population by splitting one random sample into corresponding subsets (for example, males versus females). Unlike in the matched pairs designs covered in Chapter 17, there is no matching of the individuals in two-sample problems, and the two samples can be of different sizes. The computations required for inference procedures for two-sample data differ from those for matched pairs.

When comparing two groups or two conditions, these groups or conditions form, in essence, an explanatory variable (factor). You can think of two-sample problems comparing two population means as *bivariate* problems with a quantitative response variable and a two-level factor. Here are some typical two-sample problems.

EXAMPLE 18.1 Two-sample problems

a. Researchers used an animal model to examine the long-term impact of exposure to triclosan, a broad-spectrum antimicrobial agent commonly added to soaps. They randomly assigned mice to diets containing either 0.08% triclosan or no triclosan for eight months, then compared the liver weights in each group.[1] The response variable is liver weight (quantitative) and the two-level factor is triclosan exposure (0.08% or none).
b. A field biologist observes gender-based behavior in wild chimpanzees. Twelve randomly chosen young chimpanzees are tagged remotely with a dart and their behavior is monitored. The amount of time each young chimpanzee spends in contact with its mother is recorded. The biologist then compares the amount of time spent in contact with the mother by young male chimpanzees and by young female chimpanzees. The response variable is time spent with

mother (quantitative) and the two-level factor is sex of young chimpanzee (male or female).

c. What was the role of vaccination history in the California pertussis (whooping cough) epidemic of 2010? Researchers selected a random sample of 682 medical records of California children who had been diagnosed with pertussis and a control random sample of 2016 medical records of California children in the same age group who had received care from the same clinicians on the same day but were not diagnosed with pertussis. The researchers then compared the length of time since vaccination in both groups.[2] The response variable is length of time since vaccination (quantitative) and the two-level factor is pertussis diagnosis in 2010 (yes or no).

GUDKOVANDREY/Depositphotos

Comparing two populations or the population responses to two treatments starts with data analysis: Make boxplots, dotplots (for small samples), or histograms (for larger samples) and then compare the shapes, centers, and spreads of the two samples. The most common goal of inference is to compare the average or typical responses in the two populations. When data analysis suggests that both population distributions are symmetric, and especially when they are at least approximately Normal, we want to compare the population means. Here are the conditions for inference when comparing two means.

CONDITIONS FOR INFERENCE COMPARING TWO MEANS

- We have **two SRSs,** representing two distinct populations. The samples are **independent**. That is, the individuals in one sample are unrelated to the individuals in the other sample (matching violates independence, for example). We measure the same quantitative variable for both samples.

- Both populations are **Normally distributed**. The means and standard deviations of the populations are unknown. In practice, it is enough that the distributions have similar shapes and that the data have no strong outliers.

Call the variable we measure x1x_1 in the first population and x2x_2 in the second, because the variable may have different distributions in the two populations. Here is the notation we will use to describe the two populations:

Population	Variable	Population mean	Population standard deviation
1	x1x_1	μ1μ_1	σ1σ_1
2	x2x_2	μ2μ_2	σ2σ_2

There are four unknown parameters, the two means and the two standard deviations. The subscripts remind us which population a parameter describes. We want to compare the two population means, either by giving a confidence interval for their difference μ1-μ2$\mu_1 - \mu_2$ or by testing the hypothesis H0: μ1-μ2=0,$H_0 : \mu_1 - \mu_2 = 0$, which is the same as H0: μ1=μ2, $H_0 : \mu_1 = \mu_2$, a statement of no difference.

We use the sample means and standard deviations to estimate the unknown parameters. Again, subscripts remind us which sample a statistic comes from. Here is the notation that describes the samples:

Population	Sample size	Sample mean	Sample standard deviation
1	n1n_1	x-1$\bar{x}_1$	s1s_1
2	n2n_2	x-2$\bar{x}_2$	s2s_2

To perform inference about the difference μ1-μ2$\mu_1 - \mu_2$ between the means of the two populations, we start from the difference x-1-x-2$\bar{x}_1 - \bar{x}_2$ between the means of the two samples.

EXAMPLE 18.2 Daily activity and obesity

STATE: People gain weight when they take in more energy from food than they expend. James Levine and his collaborators at the Mayo Clinic investigated the link between obesity and energy spent on daily activity.[3]

Choose 20 healthy volunteers who don’t exercise. Deliberately choose 10 who are lean and 10 who are mildly obese but still healthy. Attach sensors that monitor the subjects’ every move for 10 days. Table 18.1 presents data on the time (in minutes per day) that the subjects spent standing or walking, sitting, and lying down. Do lean and obese people differ in the average time they spend standing and walking?

TABLE 18.1 Time (minutes per day) spent in three different postures by lean and obese subjects

Group	Subject	Stand/walk	Sit	Lie
Lean	1	511.100	370.300	555.500
Lean	2	607.925	374.512	450.650
Lean	3	319.212	582.138	537.362
Lean	4	584.644	357.144	489.269
Lean	5	578.869	348.994	514.081
Lean	6	543.388	585.312	506.500
Lean	7	677.188	268.188	467.700
Lean	8	555.656	322.219	567.006
Lean	9	374.831	537.031	531.431
Lean	10	504.700	528.838	396.962
Obese	11	260.244	646.281	521.044
Obese	12	464.756	456.644	514.931
Obese	13	367.138	578.662	563.300
Obese	14	413.667	463.333	532.208
Obese	15	347.375	567.556	504.931
Obese	16	416.531	567.556	448.856
Obese	17	358.650	621.262	460.550
Obese	18	267.344	646.181	509.981
Obese	19	410.631	572.769	448.706
Obese	20	426.356	591.369	412.919

PLAN: Examine the data, check the conditions for inference, and carry out a test of hypotheses. We suspect in advance that lean subjects (Group 1) are more active than obese subjects (Group 2), so we test the hypotheses

H0: μ1=μ2 (that is, μ1-μ2=0)Ha: μ1>μ2 (that is, μ1-μ2>0)

$$H_0 : \mu_1 = \mu_2 \text{ (that is, } \mu_1 - \mu_2 = 0)$$
$$H_a : \mu_1 > \mu_2 \text{ (that is, } \mu_1 - \mu_2 > 0)$$

Catchlight Visual Services/Alamy Stock Photo

SOLVE: CONDITIONS FOR INFERENCE. Are the conditions for inference met? The subjects are volunteers, so the groups are not SRSs from all lean and mildly obese adults. The study tried to recruit comparable groups: All worked in sedentary jobs, none smoked or took medication, and so on. Setting clear standards like these helps make up for the fact that we can’t reasonably get SRSs for so invasive a study. The subjects were not told that they were chosen from a larger group of volunteers because they did not exercise and were either lean or mildly obese. Because their willingness to volunteer wasn’t related to the purpose of the experiment, we will treat them as two independent SRSs.

A dotplot of the two groups (Figure 18.1) displays the data in detail. The distributions are a bit irregular, as we expect with just 10 observations. There are no clear departures from Normality such as extreme outliers or skewness. This is confirmed in a Normal quantile plot (not shown).

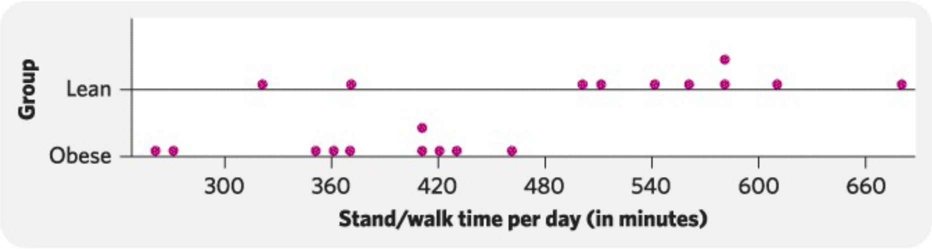


Figure 18.1

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

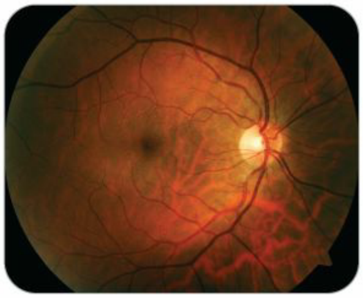
FIGURE 18.1 Dotplot of the time spent walking or standing.

To complete the Solve step, we must learn the details of inference comparing two means.

APPLY YOUR KNOWLEDGE

Which data design? *Each situation described in Exercises 18.1 to 18.4 requires inference about a mean or means. Identify the response variable and whether the setting involves (1) a single sample, (2) matched pairs, or (3) two independent samples. The procedures of Chapter 17 apply to designs (1) and (2). We are about to learn procedures for (3).*

18.1 Physiological blind spot. The physiological blind spot refers to a very small zone of functional blindness in the eye where the optic nerve passes through the retina. We do not notice it because our nervous system compensates for it. Can eye training reduce the size of a person’s physiological blind spot? Researchers recruited a representative sample of 10 adults with normal vision. Each participant performed training exercises with one eye for three weeks. The size of the physiological blind spot was measured (in degrees of visual angle squared) with a motion detection task both prior to training and again after the training was completed.[4]



esp_imaging/Getty Images

18.2 Agricultural pests. Agricultural pests can be controlled to some extent either by using pesticides to kill them or by introducing a large number of sterilized males to diminish the species' reproductive potential. Ten large corn fields are treated with pesticides, and 10 have sterilized males introduced. Corn yield (in bushels per acre) at harvest time is then compared for the two treatments.

18.3 Chemical analysis. To check a new analytical method, a chemist obtains a reference specimen of known concentration from the National Institute of Standards and Technology. She then makes 20 measurements of the concentration of this specimen with the new method and checks for bias by comparing the mean result with the known concentration.

18.4 Traumatic brain injury. People who sustain a concussion or a traumatic brain injury (TBI) are likely to experience sleep problems. Do these problems persist? Researchers recruited a sample of 31 patients who had their first ever TBI 18 months earlier and a control group of 42 otherwise similar individuals without prior brain trauma. They compared sleep duration per 24 hours (recorded electronically over a two-week period) for the two groups.[5]

TWO-SAMPLE *t* PROCEDURES

The values x-1$\bar{x}_1$ and x-2$\bar{x}_2$ are sample statistics. Therefore, the difference x-1-x-2$\bar{x}_1 - \bar{x}_2$ we actually compute is specific to the two independent random samples we selected. A different set of independent random samples would most likely yield a different value for x-1-x-2.$\bar{x}_1 - \bar{x}_2$. How much the difference x-1-x-2$\bar{x}_1 - \bar{x}_2$ can vary from sample to sample is given by its *sampling distribution.*

While the proof is beyond the scope of an introductory statistics text, it can be shown mathematically that the distribution of the difference between two independent random variables has a mean equal to the difference of their respective means and a variance equal to the sum of their respective variances. In Chapter 14 we saw that the sampling distributions of x-1$\bar{x}_1$ and x-2$\bar{x}_2$ have means μ1μ_1 and μ2μ_2 and standard deviations σ1/n1$\sigma_1/\sqrt{n_1}$ and σ2/n2,$\sigma_2/\sqrt{n_2}$, respectively. Therefore, the center of the sampling distribution of x-1-x-2$\bar{x}_1 - \bar{x}_2$ is μ1-μ2$\mu_1 - \mu_2$ (making x-1-x-2$\bar{x}_1 - \bar{x}_2$ an unbiased estimate of μ1-μ2).$\mu_1 - \mu_2$). The variance of the sampling distribution of x-1-x-2$\bar{x}_1 - \bar{x}_2$ is

σ12n1+σ22n2

$$\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

and therefore its standard deviation is

σ12n1+σ22n2

$$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

This standard deviation is smaller when the sample sizes n1n_1 and n2n_2 are increased.

When both populations are Normally distributed, the sampling distributions of x-1$\bar{x}_1$ and x-2$\bar{x}_2$ will also both be Normally distributed, and so will the sampling distribution of their difference x-1-x-2.$\bar{x}_1 - \bar{x}_2$.

Because we typically don’t know σ1σ_1 and σ2,σ_2, we estimate them by the sample standard deviations s1s_1 and s2.s_2. The result is the **standard error**, or estimated standard deviation, of the difference in sample means:

SE=s12n1+s22n2

$$\mathrm{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

When we standardize the estimate x-1-x-2,$\bar{x}_1 - \bar{x}_2$, we get

t=(x-1-x-2)-(μ1-μ2)SE

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\mathrm{SE}}$$

This **two-sample *t* statistic** has approximately a *t* distribution. It does not have exactly a *t* distribution even if the populations are both exactly Normal. In practice, however, the approximation is very accurate.

THE TWO-SAMPLE *t* STATISTIC

Draw an SRS of size n1n_1 from a large Normal population with unknown mean and standard deviation μ1μ_1 and σ1,σ_1, and draw an independent SRS of size n2n_2 from another large Normal population with unknown mean and standard deviation μ2μ_2 and σ2.σ_2. The distribution of the two-sample *t* statistic

t=(x-1-x-2)-(μ1-μ2)s12n1+s22n2

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

is very close to the *t* distribution with degrees of freedom df given by

df=s12n1+s22n221n1-1s12n12+1n2-1s22n22

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

This approximation is accurate when both sample sizes n1n_1 and n2n_2 are 5 or larger.

Unlike in the one-sample situation, the degrees of freedom of the two-sample *t* distribution depend on the variability of the sample data as well as the sample sizes. The degrees of freedom, df, are generally not a whole number. The exact value is always at least as large as the smaller of n1-1 $n_1 - 1$ and n2-1$n_2 - 1$ and at most equal to n1+n2-2.$n_1 + n_2 - 2$. There is a *t* distribution for any positive degrees of freedom, even though some statistical software and Table C at the back of the book provide degrees of freedom only as whole numbers. The impact of rounding is typically minor enough to be of no concern.

As with the one-sample situation in Chapter 17, we can use the properties of the distribution of the two-sample *t* statistic to obtain a confidence interval about the unknown parameter value μ1-μ2$\mu_1 - \mu_2$ or to test a set of null and alternative hypotheses about the unknown parameter value μ1-μ2.$\mu_1 - \mu_2$. The logic and interpretation are similar.

THE TWO-SAMPLE *t* PROCEDURES

Draw an SRS of size n1n_1 from a large Normal population with unknown mean μ1,μ_1, and draw an independent SRS of size n2 n_2 from another large Normal population with unknown mean μ2,μ_2, for the purpose of inference about the unknown quantity μ1-μ2.$\mu_1 - \mu_2$.

A level CC **confidence interval for** μ1-μ2$\boldsymbol{\mu_1 - \mu_2}$ is given by

(x-1-x-2)±t*s12n1+s22n2

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t*t^* is the critical value with area CC between -t*$-t^*$ and t*t^* under the *t* density curve with the appropriate degrees of freedom.

To **test the hypothesis** H0: μ1-μ2=0$H_0: \mu_1 - \mu_2 = 0$ (equivalent to μ1=μ2),$\mu_1 = \mu_2$), obtain the **two-sample *t* statistic**

t=x-1-x-2s12n1+s22n2

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where the test ***P*-value** is the probability, when H0H_0 is true, of getting a test statistic *t* at least as extreme in the direction of Ha H_a as that obtained, and is computed as the corresponding area under the *t* distribution with the appropriate degrees of freedom.

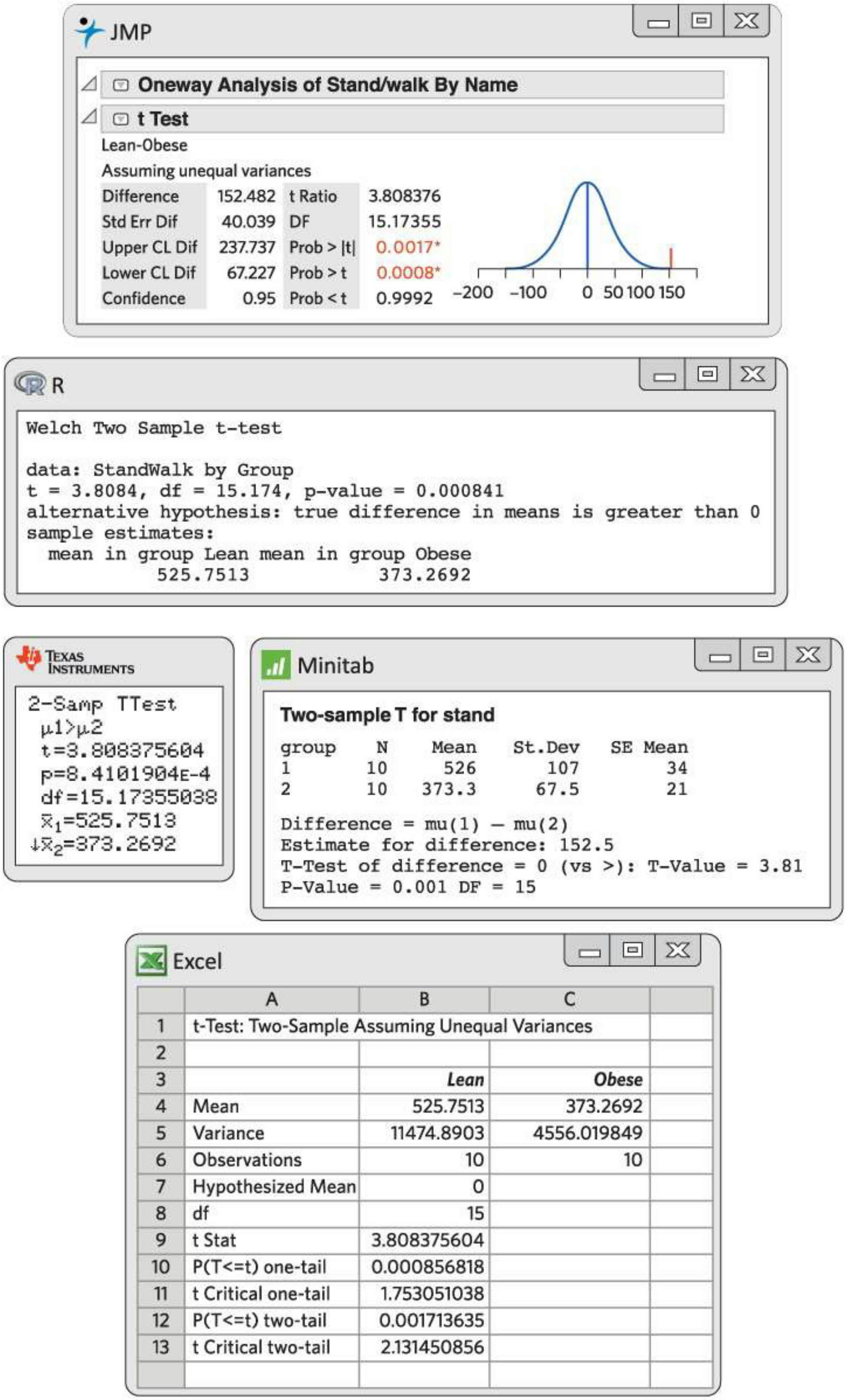
Any technology with basic statistical inference capabilities will perform all necessary computations and produce the lower and upper boundaries for a confidence interval or the test statistic and the *P*-value for a hypothesis test.

These procedures are accurate when both populations are Normal and both sample sizes n1n_1 and n2n_2 are 5 or larger. They are approximately correct for large enough sample sizes otherwise.

EXAMPLE 18.3 Daily activity and obesity

We can now complete Example 18.2.

SOLVE: INFERENCE. Software will find the summary statistics for both samples, the test statistic *t*, and the *P*-value, as shown in Figure 18.2. The sample means, in particular, are x-1=525.75 $\bar{x}_1 = 525.75$ and x-2=373.2$\bar{x}_2 = 373.2$ minutes, for a difference x-1-x-2=152.48$\bar{x}_1 - \bar{x}_2 = 152.48$ minutes. The test statistic is t=3.81,$t = 3.81$, rounded to two decimal places. The one-sided *P*-value is reported as 0.0008 (Minitab rounds all *P*-values to three decimal places, so it gives P=0.001$P = 0.001$).



Oneway Analysis of Stand/walk By Name

t Test

Lean-Obese

Assuming unequal variances

Difference	152.482	t Ratio	3.808376
Std Err Dif	40.039	DF	15.17355
Upper CL Dif	237.737	Prob > |t|	0.0017*
Lower CL Dif	67.227	Prob > t	0.0008*
Confidence	0.95	Prob < t	0.9992

```
Welch Two Sample t-test

data: StandWalk by Group
t = 3.8084, df = 15.174, p-value = 0.000841
alternative hypothesis: true difference in means is greater than 0
sample estimates:
  mean in group Lean mean in group Obese
          525.7513            373.2692
```

```
2-Samp TTest
 μ1>μ2
 t=3.808375604
 p=8.4101904E-4
 df=15.17355038
 x̄1=525.7513
↓x̄2=373.2692
```

Two-sample T for stand

```
group    N    Mean    St.Dev   SE Mean
1       10     526      107        34
2       10   373.3     67.5        21

Difference = mu(1) − mu(2)
Estimate for difference: 152.5
T-Test of difference = 0 (vs >): T-Value = 3.81
P-Value = 0.001 DF = 15
```

	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		*Lean*	*Obese*
4	Mean	525.7513	373.2692
5	Variance	11474.8903	4556.019849
6	Observations	10	10
7	Hypothesized Mean	0	
8	df	15	
9	t Stat	3.808375604	
10	P(T<=t) one-tail	0.000856818	
11	t Critical one-tail	1.753051038	
12	P(T<=t) two-tail	0.001713635	
13	t Critical two-tail	2.131450856	

Figure 18.2

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 18.2 The two-sample *t* procedures applied to the data on activity and obesity:

]

output from a graphing calculator, three statistical programs, and a spreadsheet program.

CONCLUDE: There is very strong evidence (P=0.0008)$(P = 0.0008)$ that, on average, lean people spend more time walking and standing than do moderately obese people. In the study, the average daily walking and standing times were roughly 526 minutes for the lean group and 373 minutes for the moderately obese group.

The computations performed by software in this example involve the following steps:

- Summary statistics:

Group	*n*	**Mean** x-$\bar{x}$	**Std. dev.** *s*
Group 1 (lean)	10	525.751	107.121
Group 2 (obese)	10	373.269	67.498

- Degrees of freedom of the *t* distribution:

df=107.121210+67.498210219107.1212102+1967.4982102=1147.491

$$\mathrm{df} = \frac{\left(\frac{107.121^2}{10}+\frac{67.498^2}{10}\right)^2}{\frac{1}{9}\left(\frac{107.121^2}{10}\right)^2+\frac{1}{9}\left(\frac{67.498^2}{10}\right)^2} = \frac{(1147.491+455.598)^2}{\frac{1147.491^2}{9}+\frac{455.598^2}{9}}$$

$$= \frac{2{,}569{,}894}{169{,}367.2} = 15.1735 \approx 15$$

- Two-sample *t* statistic comparing the average minutes spent standing and walking in Group 1 (lean) and Group 2 (obese):

t=x-1-x-2s12n1+s22n2=525.751-373.269107.121210+67.498210=152.4821147.491+455.598=152.4824

$$
\begin{aligned}
t &= \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \\
&= \frac{525.751-373.269}{\sqrt{\frac{107.121^2}{10}+\frac{67.498^2}{10}}}=\frac{152.482}{\sqrt{1147.491+455.598}} \\
&= \frac{152.482}{40.039}=3.808
\end{aligned}
$$

- *P*-value: As shown in Figure 18.3, this is the area to the right of t=3.808$t = 3.808$ under the t(15)$t\,(15)$ distribution, because the alternative hypothesis is one-sided in the greater-than direction. The TI-83 calculator command tcdf(3.808, 1E99,15.17) gives a *P*-value of 0.00084, whereas tcdf(3.808,1E99, 15)gives a *P*-value of 0.00086.

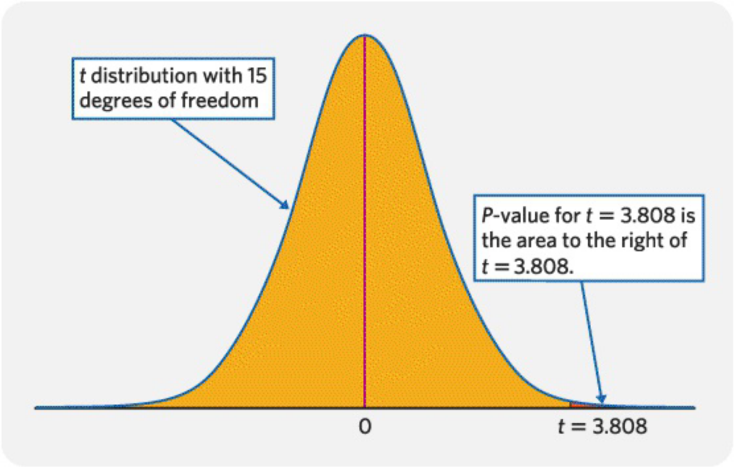


Figure 18.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 18.3 The *P*-value as the upper-tail area, for Example 18.3. We rounded down the degrees of freedom to 15.

df=15$\text{df} = 15$

t*t^*	3.733	4.073
One-sided *P*	.001	.0005

Table C at the back of the book tells us only that the one-sided *P*-value must be between 0.0005 and 0.001 because the test statistic t=3.808$t = 3.808$ is between the critical values 3.733 and 4.073 listed for the t(15)$t\,(15)$ distribution.

Software programs give accurate confidence intervals and *P*-values. However, there is some variation in software output for the two-sample *t* procedures depending on how precisely the degrees of freedom value is computed. The TI-83, R, and JMP results in Figure 18.2 are the most accurate because they use the *t* distribution with 15.17 degrees of freedom (P=0.00084).$(P = 0.00084)$. Minitab and Excel both round the degrees of freedom to a whole number. Minitab truncates the exact degrees of freedom to the next-smallest whole number, so it always errs a bit on the conservative side, whereas Excel rounds the exact degrees of freedom to the nearest whole number (in this case, both Minitab and Excel round df=15.17$\text{df} = 15.17$ to df=15$\text{df} = 15$). As a result, Excel offers P=0.00086, $P = 0.00086$, and Minitab shows P=0.001$P = 0.001$ after rounding the *P*-value to three decimal places.

These slight differences are not crucial to our conclusion, because it is the order of magnitude of the *P*-value that really matters, not its exact value. Even the more vague statement that *P* is “between 0.0005 and 0.001” from Table C is sufficient. Also, the differences seen in various software outputs get even smaller for larger sample sizes, because they have larger degrees of freedom, resulting in a proportionally smaller impact of rounding errors.

Some software outputs also show both the one-sided (“one-tail”) and two-sided (“two-tail”) *P*-values (JMP and Excel), or just the two-sided *P*-value. Remember to specify your null and alternative hypotheses before examining the sample data, and then select or calculate the one- or two-sided *P*-value that fits your alternative.

CAUTION

Note that JMP and Excel use the label “assuming unequal variances.” This label is a bit misleading: *The two-sample t procedures we have described work whether or not the two populations have the same variance.* There is a version of the two-sample *t* test we have described that works only when the two population variances are equal; we discuss this method briefly on page 462. You should favor the test that does not assume equal variances, because it will give correct results no matter what the population variances are.

Beyond the computations, it is important to always interpret the results in context and in light of the study limitations. In this example, we found a statistically significant association between the quantitative response variable (daily time spent standing and walking) and the two-level factor (body type: lean or mildly obese). *The findings come from an observational study, which means that potential confounding variables prevent us from reaching a causal conclusion.* For instance, perhaps people who gain weight are so depressed by the experience that they become less active. Or perhaps other factors, such as people's metabolism or their environment, influence both their activity level and their weight. It is impossible from this study alone to determine what causes the observed association between body type and daily activity.

Our conclusion should also reflect an understanding of the limitations of the *P*-value itself. *The P-value 0.0008 is highly statistically significant. This doesn't imply that* μleanμ_{lean} *is much larger than* μobeseμ_{obese}*, only that the evidence that* μleanμ_{lean} *is larger than* μobeseμ_{obese} *is very strong.* Obtaining a confidence interval for μlean-μobese$\mu_{lean} - \mu_{obese}$ is a better way to assess how much more lean people stand and walk, on average, compared with mildly obese people.

EXAMPLE 18.4 How much more active are lean people?

We want to estimate with 95% confidence the difference in mean daily time spent walking or standing between the two populations of lean people and mildly obese people. We already checked the conditions for inference in Example 18.2. As shown in Figure 18.2, JMP provides a 95% confidence interval for μ1-μ2$\mu_1 - \mu_2$ based on this study: "Confidence 0.95," "Lower CL Dif 67.227," and "Upper CL Dif 237.737." That is, the 95% confidence interval for μ1-μ2$\mu_1 - \mu_2$ is (67.227, 237.737).

We are 95% confident that, on average, lean people stand and walk between 67 and 238 minutes more daily than mildly obese people do.

Most of the computations required for this confidence interval were already performed for Example 18.3:

- Unbiased estimate: x-1-x-2=152.482$\bar{x}_1 - \bar{x}_2 = 152.482$
- Standard error: SE=40.039$\text{SE} = 40.039$
- Degrees of freedom: df=15.17≈15$\text{df} = 15.17 \approx 15$

The following computations are now needed:

- Critical value for a 95% confidence level: under t(15): t*=2.131 $t^* = 2.131$ (rounded to three decimal places, obtained from the Excel output “t Critical two-tail” or Table C)
- Margin of error: m=t*$m = t^*$ SE=2.131(40.039)=85.323 $\text{SE} = 2.131\,(40.039) = 85.323$
- Confidence interval, computed as estimate ±m:152.482±85.323, $\pm m : 152.482 \pm 85.323,$ or 67.159 to 237.805 minutes

Note that for df=15.17,$\text{df} = 15.17,$ t*≈2.129,$t^* \approx 2.129,$ as given by the TI-84 command invT(0.975,15.17). This explains why JMP provides the slightly narrower interval (67.227, 237.737).

The confidence interval (67, 238) is quite wide because the samples are small and the variation among individuals, as measured by the two sample standard deviations, is large. The variability of individuals is just a natural property of these populations and is beyond our control. Collecting data from substantially larger samples, however, would have provided a narrower, more informative confidence interval. Thinking about sample size early on is an important part of designing any study.

The difference in mean daily standing/walking times between the two populations is on the order of one to four hours. Any value in this range would be fairly substantial considering that the variable is a measure of daily activity. The practical relevance of such a difference in population means is a separate issue, which would need to be assessed by experts in terms of its impact on human physiology and health.

EXAMPLE 18.5 Transgenic chickens

STATE: Infection of chickens with the avian flu is a threat to both poultry production and human health. A research team created transgenic chickens resistant to avian flu infection. Could the modification affect chickens in other ways? The researchers compared the hatching weights (in grams, g) of 45 transgenic chickens and 54 independently selected commercial chickens of the same breed.[6] The data are displayed in Table 18.2.

TABLE 18.2 Hatching weights (g) of transgenic and commercial chickens

Transgenic	38.8	39.0	39.7	40.0	40.8	40.9	41.0	41.0	41.0	42.5	4
	43.0	43.4	43.5	43.5	43.8	44.4	44.7	44.7	44.7	45.3	45
	46.4	46.5	46.6	46.7	46.7	46.8	46.9	47.1	47.1	47.1	47
	47.7	48.1	48.3	49.3	49.3	49.8	50.3	50.9	52.1		
Commercial	36.7	37.1	38.9	39.5	39.5	39.8	40.0	40.2	40.3	40.5	4
	41.1	41.2	41.5	41.5	41.6	41.6	41.7	42.4	43.1	43.3	43
	43.7	44.1	44.2	45.2	45.3	45.4	46.0	46.1	46.4	46.6	46
	47.3	47.5	48.1	48.2	48.4	48.6	49.0	49.1	49.3	49.6	50
	50.4	50.6	52.2	53.0	55.5	56.4					

PLAN: We will examine the data and check the conditions for inference. We had no specific direction for the difference in weight before looking at the data, so the alternative should be two-sided. We will test the hypotheses

H0: μ1=μ2 (that is, μ1-μ2=0)Ha: μ1≠μ2 (that is, μ1-μ2≠0)

$$H_0 : \mu_1 = \mu_2 \text{ (that is, } \mu_1 - \mu_2 = 0)$$
$$H_a : \mu_1 \neq \mu_2 \text{ (that is, } \mu_1 - \mu_2 \neq 0)$$

SOLVE: The two samples can be regarded as SRSs from two populations, although the population of transgenic chickens is a potential population not yet created. Figure 18.4 shows side-by-side boxplots of the two data sets. We find no outliers (which would appear as asterisks on the graph) or skew, and we can reasonably assume a Normal distribution model for both groups. Thus, the *t* procedures should be valid.

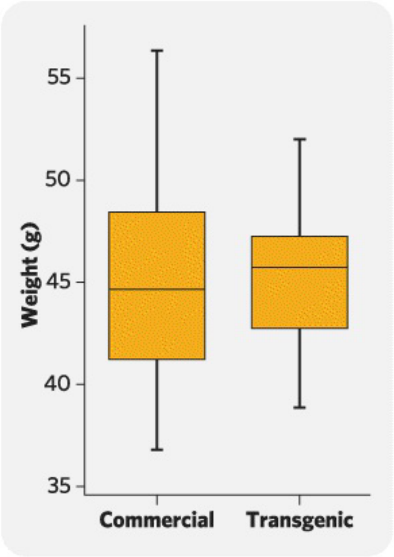
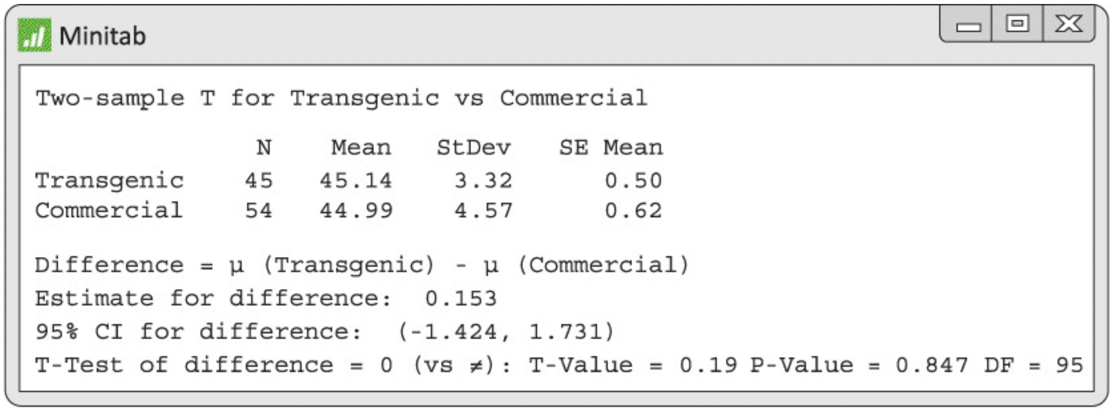


Figure 18.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 18.4 Boxplots of hatching weights of commercial and transgenic chickens.

Figure 18.5 shows the Minitab output for a two-sample *t* test with a two-sided alternative. The mean hatching weights are 45.14 g for the sample of transgenic chickens and 44.99 g for the sample of commercial chickens. The test statistic is t=0.19$t = 0.19$ and the corresponding two-sided *P*-value is 0.847, which is obviously not statistically significant.



```
Two-sample T for Transgenic vs Commercial

               N    Mean   StDev   SE Mean
Transgenic    45   45.14    3.32      0.50
Commercial    54   44.99    4.57      0.62

Difference = µ (Transgenic) - µ (Commercial)
Estimate for difference:  0.153
95% CI for difference:  (-1.424, 1.731)
T-Test of difference = 0 (vs ≠): T-Value = 0.19 P-Value = 0.847 DF = 95
```

Figure 18.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 18.5 Software output (Minitab) for chicken hatching weights.

CONCLUDE: The data give *no* significant evidence (P=0.85)$(P = 0.85)$ that the population of transgenic chickens differs in mean hatching weight from the population of commercial chickens of the same breed.

Tom Myers/AGE Fotostock

Here are the calculations involved in the two-sample *t* test:

- Summary statistics (also shown in the Minitab output):

Group	**Condition**	***n***	**x-$\bar{x}$**	***s***
1	Transgenic	45	45.14	3.32
2	Commercial	54	44.99	4.57

- Degrees of freedom of the *t* distribution:

df=3.32245+4.5725421443.322452+1534.572542=95.3

$$\text{df} = \frac{\left(\frac{3.32^2}{45}+\frac{4.57^2}{54}\right)^2}{\frac{1}{44}\left(\frac{3.32^2}{45}\right)^2+\frac{1}{53}\left(\frac{4.57^2}{54}\right)^2} = 95.3$$

- Two-sample *t* statistic:

t=x-1-x-2s12n1+s22n2=45.14-44.993.32245+4.57254=0.150.795=0.19

$$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} = \frac{45.14-44.99}{\sqrt{\frac{3.32^2}{45}+\frac{4.57^2}{54}}} = \frac{0.15}{0.795} = 0.19$$

- *P*-value: Probability, if the null hypothesis was true, of obtaining a test statistic at least as extreme in the direction of the alternative as the value computed. This is the probability under the *t*(95.3) distribution of getting a *t* value either greater than 0.19 or less than -0.19-0.19 (because our alternative is two-sided). As Figure 18.6 illustrates, the two corresponding areas have the same size so we can compute, for example, just the upper-tail area and then double it. In Excel (using a *t* value rounded to two decimal places), the command =2*T.DIST.RT(0.19,95.3) returns the value 0.84971. Minitab, which does not round *t* but rounds df to 95 gives $P = 0.847$. Again, these differences are very minor.

If no statistical software is available, Table C can be used to assess the order of magnitude of the test *P*-value. As there is no df = 95 in Table C, we select the nearest lower value available, df = 80 (it is more conservative and typically preferable to select a smaller value). Compare $t = 0.19$ with the values listed in the row for df = 80 and find the corresponding two-sided *P*. This tells us that the *P*-value for our test is greater than 0.50.

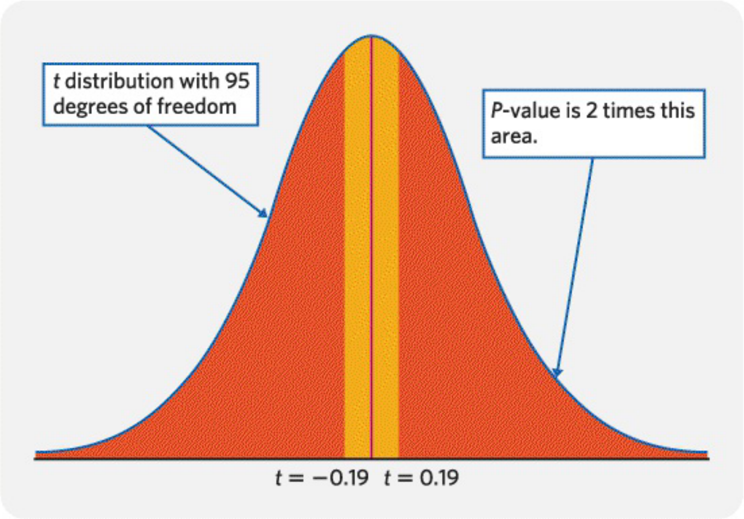


Figure 18.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 18.6 The *P*-value for Example 18.5. Because the alternative is two-sided, the *P*-value is double the area to the right of *t* = 0.19. We rounded the degrees of freedom to 95.

$df = 80$

t^*	0.678
Two-sided *P*	.50

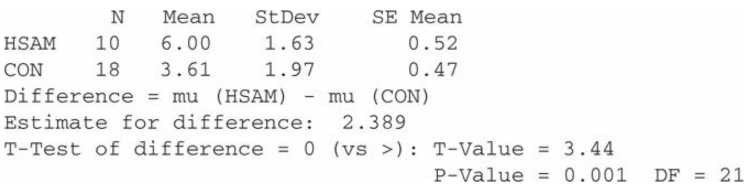
The *P*-value we found is quite large (0.85) yet this says nothing more than any non–statistically significant *P*-value. *A common mistake we discussed in Chapter 15 (page 386) is to believe that a large P-value offers strong support for the null hypothesis. A large P-value says only that the data would not be surprising if* H_0 *was* *true, but it doesn't imply that* H0H_0 *is necessarily true*. In fact, the Minitab output in Figure 18.5 reports the range -1.424-1.424 g to 1.731 g as a 95% confidence interval for μ1-μ2.$\mu_1 - \mu_2$. Therefore μ1-μ2$\mu_1 - \mu_2$ could very well be a negative value, a positive value, or even zero: Perhaps the commercial chickens are heavier at birth on average, or perhaps the transgenic chickens are heavier on average, or perhaps the genetic modification has no impact on hatching weight and the two populations have exactly the same mean hatching weight. A *P*-value that isn't small is simply inconclusive, no matter how large it is.

APPLY YOUR KNOWLEDGE

18.5 Highly Superior Autobiographical Memory Some individuals have the ability to recall accurately vast amounts of autobiographical information without mnemonic tricks or extra practice. This ability is called Highly Superior Autobiographical Memory (HSAM). A study recruited adults with confirmed HSAM and control individuals of similar age without HSAM. All study participants were given a battery of cognitive and behavioral tests with the goal of finding out how this extraordinary ability works. Here are the participants' results for a visual memory test:[7]

HSAM	4	4	5	9	6	6	7	6	8	5
Control	5	1	2	4	2	2	3	3	3	9
	4	2	4	5	5	1	6	4		

a. Following is the Minitab output for the two-sample *t* test. What are x-$\bar{x}$ and ss for each of the two samples? Starting from these values, obtain the two standard errors of the means and the *t* test statistic. Your work should agree with the output.



```
          N    Mean    StDev     SE Mean
HSAM     10    6.00     1.63        0.52
CON      18    3.61     1.97        0.47
Difference = mu (HSAM) - mu (CON)
Estimate for difference:  2.389
T-Test of difference = 0 (vs >): T-Value = 3.44
                                  P-Value = 0.001   DF = 21
```

b. Do individuals with HSAM have significantly greater abilities on visual memory tests than typical individuals? Summarize the findings in a sentence or two, including *t*, df, *P*, and a conclusion, as if you were preparing a report for publication.

18.6 Traumatic brain injury, continued Go back to Exercise 18.4, which compared 31 patients who had suffered a TBI 18 months earlier with a control group of 42 individuals with no prior brain trauma. Here is a finding reported in a scientific journal: “In the chronic state after traumatic brain injury, sleep need per 24 hours was persistently increased in trauma patients (8.1±0.5$(8.1 \pm 0.5$ hours) as compared to healthy controls (7.1±0.7$(7.1 \pm 0.7$ hours).” The values cited are group mean plus or minus standard deviation.

a. There searchers performed a two-sided two-sample *t* test and reported a highly statistically significant *P*-value (<0.00005). (< 0.00005). Compute the test statistic using the cited summary values. Obtain the test *P*-value from software or Table C (df=70.94).$(\mathrm{df} = 70.94)$.
b. Obtain a 95% confidence interval for µ1-µ2,$\mu_1 - \mu_2$, the difference in mean sleep need per 24 hours for the population of subjects with a past TBI and the control, trauma-free population.
c. Write a summary of the findings in context.

18.7 Echinacea for the common cold? Echinacea is widely used as an herbal remedy for the common cold, but does it work? In a double-blind experiment, healthy volunteers agreed to be exposed to common–cold–causing rhinovirus type 39 and have their symptoms monitored. The volunteers were randomly assigned to take either a placebo or an echinacea supplement daily from seven days before until five days after viral exposure. A symptom score was recorded for each subject over the five days following exposure, with higher scores indicating more severe symptoms. The published results reported the mean±SEM$\mathrm{mean} \pm \mathrm{SEM}$ for both groups as 13.21±1.91 13.21 ± 1.91 (echinacea) and 15.05±1.43$15.05 \pm 1.43$ (placebo).[8]

a. The two-sample *t* statistic for x-1-x-2$\bar{x}_1 - \bar{x}_2$ was t=-0.771. $t = -0.771$. You can draw a conclusion from this *t* without using a table and even without knowing the sizes of the samples (remember to specify your null and alternative hypotheses first). What is your conclusion? Why don’t you need the sample sizes and a table?

b. In all, 52 subjects were assigned to the echinacea treatment and 103 to the placebo. Fill in the values in this summary table:

Group	Treatment	*n*	x-$\bar{x}$	*s*
1	Echinacea	?	?	?
2	Placebo	?	?	?

Compute the degrees of freedom and use software or Table C to obtain the *P*-value.

18.8 What color is that dress? In 2015, people all over the internet were hotly debating whether a specific dress was actually white and gold (WG) or black and blue (BB). The photograph was unusually lit and provided little context beyond the dress itself, leaving much to interpretation. The dress therefore made an excellent tool to study the involvement of higher cognitive functions in color perception. Researchers recruited a random sample of adults, some of whom perceived the dress to be WG and some of whom perceived it to be BB. While the subjects were looking at the picture of the dress on a computer screen, their relative brain activity in the frontal cortex was assessed using functional magnetic resonance imaging (fMRI) brain imaging. Here are the findings:[9]

BB	− 0.68	− 0.58	− 0.57	− 0.42	− 0.39	− 0.36	− 0.30
	− 0.21	− 0.23	− 0.14	− 0.12	− 0.08	0.04	0.20
WG	− 0.24	− 0.03	0.07	0.15	0.16	0.17	0.33
	0.35	0.37	0.49	0.52	0.54	0.68	0.76

Do the data give significant evidence that, on average, brain activity in the frontal cortex differs based on how the dress is perceived? Use the four-step process to answer this question.

Daniel Karmann/picture-alliance/
dpa/AP Images

ROBUSTNESS AGAIN

The two-sample *t* procedures are more robust than the one-sample *t* methods, particularly when the distributions are not symmetric. When the sizes of the two samples are equal and the two populations being compared have distributions with similar shapes, probability values from the *t* density curve are quite accurate for a broad range of population distributions when the sample sizes are as small as n1=n2=5.$n_1 = n_2 = 5$.[10] When the two populations have different shapes, larger samples are needed.

As a guide to practice, we can adapt the guidelines given on page 435 for the use of one-sample *t* procedures to two-sample procedures by replacing "sample size" with "sum of the sample sizes," n1+n2.$n_1 + n_2$. These guidelines err on the side of safety, especially when the two samples are of equal size. *In planning a two-sample study, choose equal sample sizes whenever possible. The two-sample t procedures are most robust against non-Normality in this case.*

CAUTION

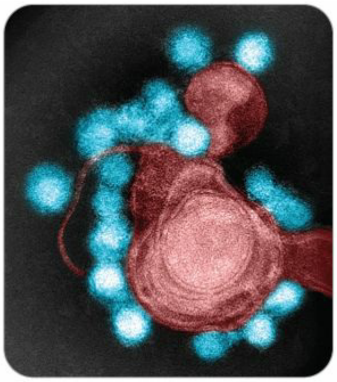
APPLY YOUR KNOWLEDGE

18.9 Zika and male reproduction. The Zika virus, which is known to be transmitted by mosquitoes, can also be transmitted sexually, as the virus can persist for weeks in the semen of symptom-free infected men. Using a mouse model, researchers examined the impact of Zika on male reproductive organs. Young adult male lab mice of a breed susceptible to Zika were either inoculated with Zika or not. Here are the weights of their testes two weeks later (in grams, g):[11]

Uninfected	85	85	90	91	96	100		
Infected	17	93	15	12	10	9	10	12

a. What two populations do the researchers want to compare and what is the response variable? State the null and alternative hypotheses.

b. Plot the data and check the conditions for inference. Explain why a two-sample *t* test is not appropriate here.



Science Source

18.10 Reducing snoring. Snoring is an extremely common and burdensome condition that occurs when vibrating soft tissue obstructs the upper airway during sleep. Researchers investigated the effectiveness of throat exercises designed to strengthen that soft tissue. They recruited otherwise healthy adult patients who had recently sought medical care for a primary complaint of snoring. At random, the subjects were divided into two treatment groups, with the treatments lasting for three months. The control group was asked to perform deep breathing exercises daily and to sleep with a nasal dilator strip. The therapy group was instructed to perform a series of throat exercises three times a day. After these three months, subjects were recorded during sleep to assess their snoring frequency (in number of snores per hour). Here are the findings:[12]

Control	2	4	6	6	6	18	23	46	55	75
	139	174	198	218	223	294	420	527	540	637
Treatment	7	10	18	27	51	52	53	69	87	101
	104	141	155	182	222	256	389	416	626	

Does the study provide significant evidence that throat exercises are better than nasal strips and breathing exercises in addressing snoring in this population? Follow the four-step process, making sure to check the conditions for inference.

AVOID THE POOLED TWO-SAMPLE *t* PROCEDURES[i]

Most software packages, including those illustrated in Figure 18.5, offer a choice of two-sample *t* statistics. One is often labeled for “unequal” variances, and the other for “equal” variances. The “unequal” variance procedure is our two-sample *t*. *This test is valid whether or not the population variances are equal.* The other choice is a special version of the two-sample *t* statistic that assumes the two populations have the same variance. It averages (the statistical term is “pools”) the two sample variances to estimate the common population variance. The resulting statistic is called the *pooled two-sample t statistic.* It is calculated as

tpooled=x-1-x-2spooled1n1+1n2=x-1-x-2(n1-1)s12+(n2-1)s22n1+n2-21n1+1n2

$$t_{pooled}=\frac{\bar{x}_1-\bar{x}_2}{s_{pooled}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

where spooled2s^2_{pooled} is the pooled sample variance. The pooled *t* statistic (which is directly related to the ANOVA procedure we will discuss in Chapter 24) is equal to our *t* statistic if the two sample sizes are the same, but not otherwise.

We could choose to use the pooled *t* for tests and confidence intervals. The pooled *t* statistic has exactly the *t* distribution with n1+n2-2n_1+n_2-2 degrees of freedom *if* the two population variances really are equal and the population distributions are exactly Normal. Of course, in the real world, distributions are not exactly Normal, and population variances are not exactly equal. The requirement for equal population variances is one more test assumption to check, and simulations show that the “unequal” variances *t* procedures are almost always more accurate than the pooled procedures.[13] Our advice: *Always use the t procedures for “unequal” variances if you can.*

STATS IN YOUR WORLD

META-ANALYSIS

Small samples have large margins of error. Large samples are expensive. Often we can find several studies of the same issue; if we could combine their results, we would have a large sample with a small margin of error. That is the idea behind “meta-analysis.” Of course, we can’t just lump the studies together, because of differences in their design and quality. Statisticians have more sophisticated ways of combining the results. Meta-analysis has been applied to issues ranging from the effects of secondhand smoke to a genetic basis of cognitive empathy.

AVOID INFERENCE ABOUT STANDARD DEVIATIONS[i]

In a two-sample setting with quantitative data, we may wish to compare either the *centers* or the *spreads* of the two groups. In a Normal population, we measure the center by the mean and the spread by the standard deviation. We use the t procedures for inference about population means for Normal populations, and we know that t procedures are widely useful for non-Normal populations as well. It is natural to turn next to inference about the standard deviations of Normal populations. Our advice here is short and clear: *Don't do it without expert advice.*

There are methods for inference about the standard deviations of Normal populations. The most common such method is the FF test for comparing the spread of two Normal populations. *Unlike the t procedures for means, the* FF *test is extremely sensitive to non-Normal distributions.* This lack of robustness does not improve in large samples. It is difficult in practice to tell whether a significant test result is evidence of unequal population spreads or simply a sign that the populations are not Normal.

The deeper difficulty underlying the very poor robustness of Normal population procedures for inference about spread has already been seen in our work on describing data. The standard deviation is a natural measure of spread for Normal distributions, but not for distributions in general. In fact, because skewed distributions have unequally spread tails, no single numerical measure does a good job of describing the spread of a skewed distribution. In summary, the standard deviation is not always a useful parameter, and even when it is (for symmetric distributions), the results of inference are not always trustworthy. Consequently, *we do not recommend trying to perform inference about population standard deviations in basic statistical practice.*[14]

CHAPTER 18 SUMMARY

- The data in a **two-sample problem** are two independent SRSs, each drawn from a separate population.
- Tests and confidence intervals for the difference μ1-μ2$\mu_1 - \mu_2$ between the means of two Normal populations start from the difference x-1-x-2 $\bar{x}_1 - \bar{x}_2$ between the two sample means.
- Draw independent SRSs of sizes n1n_1 and n2n_2 from two Normal populations with parameters μ1,μ_1, σ1σ_1 and μ2,μ_2, σ2.σ_2. The **two-sample *t* statistic** is

t=(x-1-x-2)-(μ1-μ2)s12n1+s22n2

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The statistic *t* has approximately a *t* distribution with **degrees of freedom**

df=s12n1+s22n221n1-1s12n12+1n2-1s22n22

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

In practice, use technology to obtain the degrees of freedom and accurate probability values.

- A **confidence interval for** μ1-μ2$\mu_1 - \mu_2$ is

(x-1-x-2)±t*s12n1+s22n2

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The critical value t*t^* for the computed degrees of freedom gives a confidence level very close to the desired level C.C.

- A **hypothesis test for** H0: μ1=μ2$H_0 : \mu_1 = \mu_2$ is based on the test statistic

  t=x-1-x-2s12n1+s22n2

  $$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

  The area under the *t* distribution with the computed degrees of freedom corresponding to values at least as extreme (in the direction of the alternative) as the computed test statistic provides a very accurate *P*-value.
- Because of the central limit theorem, the two-sample *t* procedures are approximately correct for non-Normal population distributions when the sample sizes are large enough. Equal sample sizes are recommended for optimal robustness against the assumption of Normality.
- The guidelines for practical use of two-sample *t* procedures and their limitations (such as what the margin of error doesn't cover or what a *P*-value doesn't say) are similar to those for one-sample *t* procedures.

THIS CHAPTER IN CONTEXT

In Chapter 17 we described the one-sample *t* procedures used to perform statistical inference for a population mean μ.μ. We also saw that this method can be applied to matched pairs designs when we are interested in the population mean difference μμ of paired observations. Chapter 18 extends the use of the *t* procedures to compare the means, μ1μ_1 and μ2,μ_2, of two distinct populations. In Chapter 24 we will see how another method (ANOVA) can be used to compare the means from several populations.

Which procedure should be used depends on how the data were collected; we discussed data collection issues in Chapters 6 and 7. Recall that conclusions from observational studies and experiments are substantially different because observational studies have confounding variables that prevent us from assigning causality. Therefore statistical significance should always be interpreted in the context of a study’s design, strengths, and weaknesses.

While Chapters 17 and 18 have dealt with inference for a quantitative variable (with the population mean μμ as the key parameter), the next two chapters will introduce inference procedures used for a categorical variable (with the population proportion pp as the key parameter). You will notice that Chapters 19 and 20 closely mirror Chapters 17 and 18 in their approach, first addressing one-sample procedures then moving on to two-sample procedures. Often what students find the most challenging is figuring out whether a variable is quantitative or categorical, a topic that we discussed in great detail in Chapter 1.

CHECK YOUR SKILLS

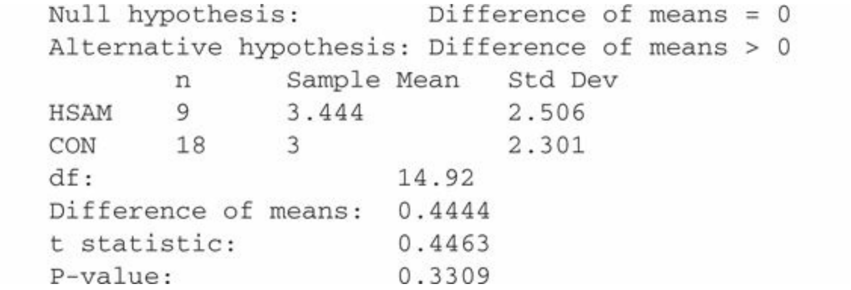
18.11 Topiramate can reduce drinking in alcohol-dependent individuals but would it work for heavy drinkers who are not alcohol-dependent? A study recruited 138 heavy drinkers who wanted to reduce their alcohol consumption and randomly assigned them to take either topiramate or a placebo for three months. The study report states that "topiramate treatment significantly reduced heavy drinking days relative to placebo."[15] Which procedure is used to form this conclusion?

a. The one-sample *t* test
b. The matched pairs *t* test
c. The two-sample *t* test

18.12 Do chimpanzees trust their friend? Using a game setting, researchers were able to compute a trust index based on how often the main player would enlist the help of a partner, hoping that the partner would share the reward. Each chimpanzee performed the game, in random order, once when interacting with a close friend chimpanzee and once when interacting with a non-friend chimpanzee. The study report states that "chimpanzees trust their friends significantly more frequently than their non-friends."[16] Which procedure is used to form this conclusion?

a. The one-sample *t* test
b. The matched pairs *t* test
c. The two-sample *t* test

Use the following information for Exercises 18.13 and 18.14. The study in Exercise 18.5 compared individuals who have Highly Superior Autobiographical Memory (HSAM) with control individuals without this ability. Participants were asked to take a verbal association test. Here is the CrunchIt! output for the two-sample t test asking whether individuals *with HSAM score significantly higher than control individuals on such a measure of verbal association.*



```
Null hypothesis:        Difference of means = 0
Alternative hypothesis: Difference of means > 0
        n       Sample Mean   Std Dev
HSAM    9       3.444         2.506
CON     18      3             2.301
df:                    14.92
Difference of means:   0.4444
t statistic:           0.4463
P-value:               0.3309
```

18.13 What is the value of the test statistic for this two-sample *t* test?

a. 0
b. 0.4463
c. 12.92

18.14 From the two-sample *t* test we conclude that

a. individuals with HSAM have significantly greater abilities on verbal association tests than non-HSAM individuals, on average.
b. individuals with or without HSAM have the same abilities on verbal association tests, on average.
c. the study failed to find significant evidence that, on average, individuals with HSAM have greater abilities on verbal association tests than non-HSAM individuals.

Use the following information for Exercises 18.15 to 18.18. A study of the effects of exercise used rats bred to have high or low capacity for exercise. The 8 high-capacity rats had mean blood pressure 89 millimeters of mercury (mm Hg) and standard deviation 9 mm Hg; the 8 low-capacity rats had mean blood pressure 105 mm Hg with standard deviation 13 mm Hg.

18.15 What is the value of the two-sample *t* statistic for comparing the two population means?

a. 0.5

b. 2.86
c. 9.65

18.16 We suspect that rats with a low capacity for exercise tend to have a higher blood pressure than rats with a high capacity for exercise. To see if this is true, test these hypotheses for the mean blood pressure of all rats with high capacity (HC) and all rats with low capacity (LC) for exercise:

a. H0: μHC=89;μLC=105$H_0: \mu HC = 89; \mu_{LC} = 105$ versus Ha: μHC≠89;μLC≠105$H_a: \mu_{HC} \neq 89; \mu_{LC} \neq 105$
b. H0: μHC=μLC$H_0: \mu_{HC} = \mu_{LC}$ versus Ha: μHC≠μLC$H_a: \mu_{HC} \neq \mu_{LC}$
c. H0: μHC=μLC$H_0: \mu_{HC} = \mu_{LC}$ versus Ha: μHC<μLC$H_a: \mu_{HC} < \mu_{LC}$

18.17 The two-sample *t* test has degrees of freedom df = 12.46. The *P*-value corresponding to the hypotheses specified in Exercise 18.16 satisfies

a. 0.005<P<0.01.$0.005 < P < 0.01.$
b. 0.01<P<0.02.$0.01 < P < 0.02.$
c. 0.02<P<0.05.$0.02 < P < 0.05.$

18.18 We are 95% confident that the difference μLC-μHC$\mu_{LC} - \mu_{HC}$ is

a. 16 mm Hg.
b. a value between 3.9 and 28.1 mm Hg.
c. a value between 5.1 and 26.9 mm Hg.

18.19 One major reason that the two-sample *t* procedures are widely used is that they are quite *robust*. This means that

a. *t* procedures do not require that we know the standard deviations of the populations.
b. confidence levels and *P*-values from the *t* procedures are quite accurate even if the population distribution is not exactly Normal.
c. *t* procedures compare population means—a comparison that answers many practical questions.

18.20 A study of road rage asked samples of 596 men and 523 women about their behavior while driving. Based on their answers, each subject was assigned a road rage score on a scale of 0 to 20. The subjects were chosen by random digit dialing of telephone numbers. Are the conditions for two-sample *t* inference satisfied?

a. Maybe: The SRS condition is acceptable but we need to look at the data to check Normality.

b. No: Scores in a range between 0 and 20 can’t be Normal.

c. Yes: The SRS condition is acceptable and large sample sizes make the Normality condition unnecessary.

CHAPTER 18 EXERCISES

In exercises that call for two-sample t procedures, either use technology or do your calculations by hand and use Table C after rounding the degrees of freedom down. Many of the following exercises ask you to think about issues of statistical practice as well as to carry out procedures.

18.21 Sexual dimorphism in birds. The central tail feathers of the long-tailed finch (*Poephila acuticauda*) are a sexually dimorphic trait hypothesized to play a role in sexual selection. The longer tail feathers in males cost energy to produce, and this is thought to signal the male's excellent health condition. Here are the lengths of the central tail feathers (average of the two central feathers, in millimeters) of 20 male and 21 female long-tailed finches:[17]

Males									
87	77	95	73	74	85	56	86	95	108
75	87	73.5	82	89	64	74.5	87	85	86
Females									
60	59	72	54	65	58	59	60	65	60
68	70.5	80	87	65	59	65	67	62	66
70									

a. Treat these data as SRSs from the population of adult long-tailed finches in the wild. Make dotplots of both data sets and determine whether the use of a two-sample *t* procedure is appropriate.

b. How much longer, on average, are the central tail feathers of male long-tailed finches than those of the females? Give a 95% confidence interval for the difference in population mean length between the male and the female adult long-tailed finches.

18.22 Food counseling for obese children. According to the Centers for Disease Control and Prevention (CDC), obesity rates among U.S. children have increased dramatically over the past three decades, from a low of about 5% to 18% as of 2010. A study examined the effect of family-based food-counseling sessions provided by trained professionals. The study randomly assigned obese children aged 9 to 12 years to either the counseling intervention or a control group not receiving any food counseling. The children’s weight changes (in pounds, lb) after 15 weeks are displayed in Table 18.3 for both conditions.[18]

a. Plot both data sets using dotplots or histograms. Because the sample sizes are quite large, the *t* procedures will work well even in the presence of skew or mild outliers.
b. Is there good evidence that obese children receiving food counseling gain less weight over a 15-week period? State the hypotheses, carry out the corresponding test, and conclude in context.

TABLE 18.3 Weight change (lb) of obese children, by group

Intervention									
− 16.7	− 14.8	− 11.9	− 9.7	− 9.6	− 8.8	− 8.0	− 7.1	− 6.6	− 6.0
−5.6	−5.6	−5.5	−5.5	−5.1	−5.0	−5.0	−4.8	−4.4	−4.4
−4.1	−4.0	−4.0	−3.6	−3.5	−3.2	−2.8	−2.0	−1.8	−1.8
−1.4	−1.2	−0.2	−0.1	0.0	0.2	0.6	1.0	1.2	1.2
1.4	1.8	2.0	2.2	2.5	2.8	3.3	4.2	5.4	5.8
6.0	6.4	8.4							
Control									
120	10.0	9.0	8.6	7.1	6.7	6.1	4.8	4.6	4.5
0.1	−5.1	−5.1							

18.23 IQ scores for boys and girls. Here are the IQ test scores of 31 seventh-grade girls in a midwestern school district:[19]

114	100	104	89	102	91	114	114	103	105	
108	130	120	132	111	128	118	119	86	72	
111	103	74	112	107	103	98	96	112	112	93

The IQ test scores of 47 seventh-grade boys in the same district are as follows:

111	107	100	107	115	111	97	112	104	106	113
109	113	128	128	118	113	124	127	136	106	123
124	126	116	127	119	97	102	110	120	103	115
93	123	79	119	110	110	107	105	105	110	77
90	114	106								

a. Make dotplots or histograms of both sets of data. Because the distributions are reasonably symmetric with no extreme outliers, the *t* procedures will work well.

b. Treat these data as SRSs from all seventh-grade students in the district. Is there good evidence that girls and boys differ in their mean IQ scores?

18.24 Food counseling for obese children, continued.

a. Use the data in Exercise 18.22 to give a 95% confidence interval for the difference in mean weight change after 15 weeks for obese children who receive professional food counseling and obese children who do not.

b. Give a 95% confidence interval for the mean weight change after 15 weeks for obese children who receive professional food counseling.

18.25 Bird songs. Bird songs have been hypothesized to be a secondary sexual character signaling an individual’s health status. Researchers designed an experiment in which they randomly assigned male collared flycatchers (*Ficedula albicollis*) to two groups: One group received an immune challenge in the form of an injection of sheep red blood cells, and the other group received a placebo injection. Here are the changes in song rate (in strophes per minute) after the injection for 15 male collared flycatchers in the immune-challenge group and 12 males in the placebo group:[20]

Immune challenge

−1.6	−3.1	−2.7	−3.7	−3.1	−3.6	−1.9	−1.5
−0.1	0.8	−0.1	−0.2	−1.2	−1.9	0.2	
Placebo							
−1.5	1.7	0.4	−1.8	0.0	0.4	0.8	2.0
0.0	−2.4	−1.5	−0.1				

a. Make dotplots to investigate the shape of the distributions. Is the use of a two-sample *t* procedure appropriate?

b. Do the data provide significant evidence that an immune challenge reduces the male song rate, on average, more than a placebo injection does? Obtain the test statistic, degrees of freedom, and *P*-value, and state your conclusion. Does the study support the hypothesis that male bird songs advertise the male's health status?

18.26 Testosterone and obesity in adolescent males. Obesity in adult males is associated with lower levels of sex hormone. Example 17.2 described a study investigating a possible link between obesity and testosterone levels in adolescent males between the ages of 14 and 20 years. The study compared 25 obese adolescent males and 25 adolescent males with a healthy weight. Here are their plasma testosterone concentrations (in nanomoles per liter, nmol/l):

Obese adolescent males								
0.30	0.24	0.19	0.17	0.18	0.23	0.24	0.06	0.15
0.17	0.18	0.17	0.15	0.12	0.25	0.25	0.25	0.32
0.35	0.37	0.39	0.46	0.49	0.42	0.36		
Adolescent males with a healthy weight								
0.78	0.70	0.63	0.60	0.60	0.69	0.76	0.58	0.50
0.48	0.49	0.43	0.42	0.38	0.35	0.35	0.32	0.31
0.28	0.25	0.23	0.24	0.24	0.26	0.27		

a. Plot each data set to investigate the shapes of the distributions. Are the data appropriate for a *t* procedure?

b. We suspect that obese adolescent males have a lower plasma testosterone level, on average. Do the data support this suspicion? Carry out the appropriate test and state your conclusion.

c. Give a 95% confidence interval for the difference in mean plasma testosterone level between adolescent males who are obese and those who have a healthy weight.

18.27 Diabetic mice. The body’s natural electrical field helps wounds heal. If diabetes changes this field, that might explain why people with diabetes heal slowly. A study of this idea compared normal mice and mice bred to spontaneously develop diabetes. The investigators attached sensors to the right hip and front feet of the mice and measured the difference in electrical potential (millivolts) between these locations. Here are the data:[21]

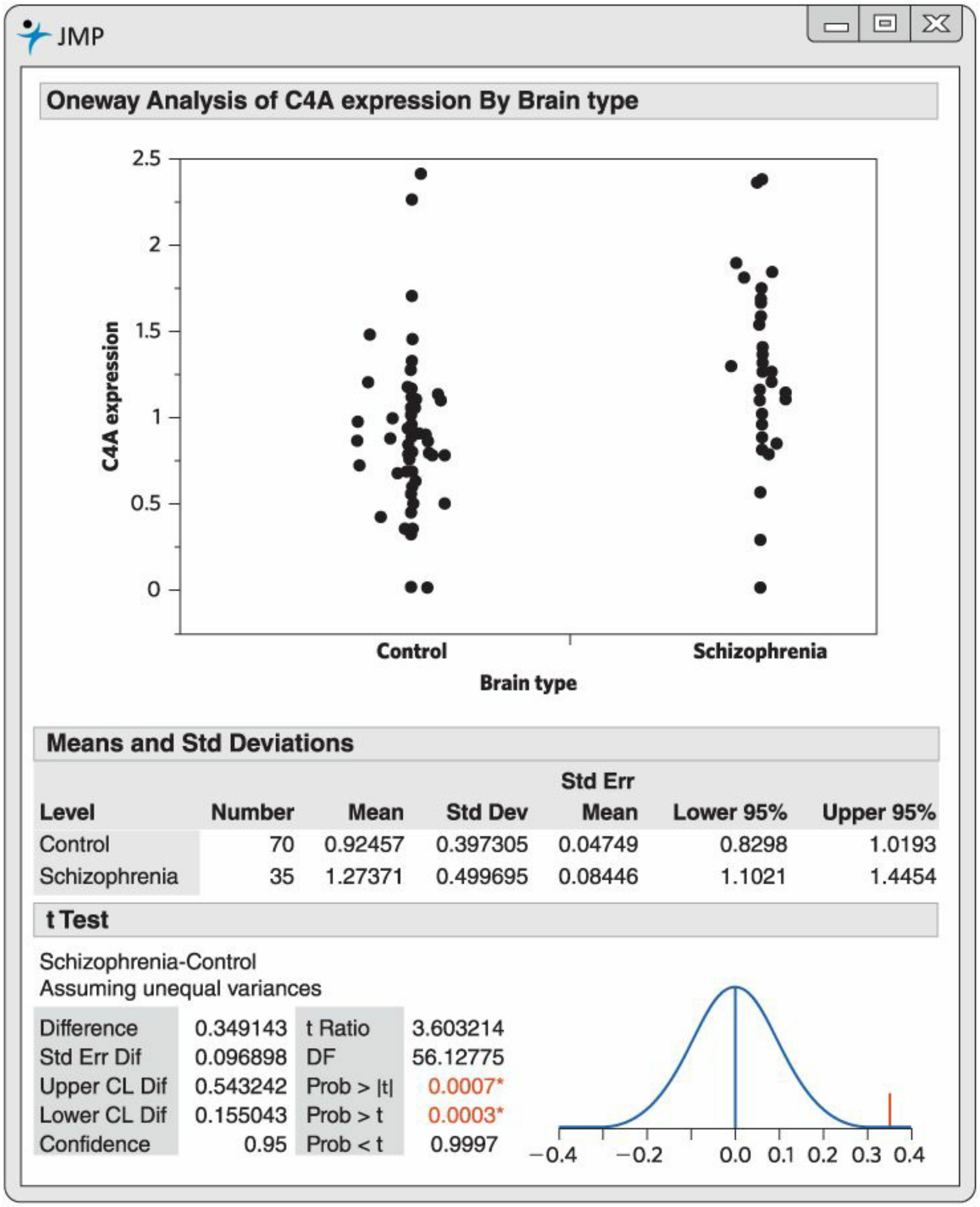
Diabetic mice					Normal mice			
13.60	7.40	1.05	10.55	16.40	13.80	9.10	4.95	7.70
22.60	15.20	19.60	17.25	18.40	7.20	10.00	14.55	13.30
11.70	14.85	14.45	18.25	10.15	9.50	10.40	7.75	8.70
10.30	10.45	8.55	8.85	19.20	8.40	8.55	12.60	9.40
10.00	9.80	10.85	14.70		6.65	8.85		

a. Make a dotplot of each sample of potentials. There is a low outlier in the diabetic group. Does it appear that potentials in the two groups differ in a systematic way?

b. Is there significant evidence of a difference in mean potentials between the two groups?

c. Repeat your inference without the outlier. Does the outlier affect your conclusion?

18.28 Schizophrenia. Schizophrenia is a heritable mental disorder with a complex genetic component in which multiple genes appear to play a part in the disorder’s emergence. A study examined the involvement of the gene variant C4A within the major histocompatibility complex in the immune system. The researchers measured the expression level of C4A (from its RNA) in brains donated to science post-mortem by 35 patients diagnosed with schizophrenia and 70 individuals not affected by schizophrenia. Figure 18.7 displays the full JMP output for a two-sample *t* test on the published data.[22]



Oneway Analysis of C4A expression By Brain type

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
Control	70	0.92457	0.397305	0.04749	0.8298	1.0193
Schizophrenia	35	1.27371	0.499695	0.08446	1.1021	1.4454

t Test

```
Schizophrenia-Control
Assuming unequal variances
Difference      0.349143   t Ratio     3.603214
Std Err Dif     0.096898   DF          56.12775
Upper CL Dif    0.543242   Prob > |t|   0.0007*
Lower CL Dif    0.155043   Prob > t     0.0003*
Confidence          0.95   Prob < t     0.9997
```

Figure 18.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 18.7 Software output (JMP) for C4A expression levels in patients with and without schizophrenia.

a. Assume that the brains donated to science are representative of the larger populations of individuals with or without schizophrenia. Why is this two-sample *t* test appropriate?
b. Does the study provide significant evidence that C4A gene

expression is associated with schizophrenia? Conclude in context, citing values from the JMP output to support your conclusion.

c. Can we also conclude from this study that the differential expression of the C4A gene causes schizophrenia? Explain your reasoning.

Exercises 18.29 to 18.37 are based on summary statistics rather than raw data. This information is typically all that is presented in published reports. You must trust that the authors understood the conditions for inference and verified that they apply to the study data. This isn't always true.

18.29 Pharmacokinetics of a drug. Avandia (rosiglitazone maleate) is an oral antidiabetic drug produced by the pharmaceutical company GlaxoSmithKline. Before a drug can be prescribed, we must know how the body absorbs and excretes it. Patients were given a single dose of either 1 milligram (mg) or 2 mg of rosiglitazone maleate, and the maximum plasma concentration of the drug (in nanograms per milliliter, ng/ml) was assessed.[23]

Treatment	nn	x-$\bar{x}$	ss
1 mg	32	76	13
2 mg	32	156	42

Is there significant evidence that maximum plasma concentration is dose-dependent (that is, higher doses result in higher concentrations)?

18.30 More on pharmacokinetics. The study in Exercise 18.29 also looked at how fast the body gets rid of the drug Avandia by measuring the elimination half-life (in hours). Here are the results:

Treatment	nn	x-$\bar{x}$	ss
1 mg	32	3.16	0.72

2 mg	32	3.15	0.39

Is there significant evidence of a difference in elimination half-life for different drug doses? State your conclusions from this exercise and Exercise 18.29 in the context of the drug’s pharmacokinetic properties.

18.31 Tongue piercing and cracked teeth. A study compared tooth health and periodontal damage in a group of 46 young adult males with a tongue piercing and a control group of 46 young adult males without tongue piercing. One question of interest was whether individuals with tongue piercing had more enamel cracks, on average. Here are the summary statistics:[24]

Group	**n**n	**x-**$\bar{x}$	**s**s
Tongue piercing	46	4.0	3.5
No tongue piercing	46	1.2	1.3

a. Is there evidence at the 1% level that young adult males with a tongue piercing have significantly more enamel cracks? State hypotheses in terms of the two population means, obtain the two-sample *t* statistic and *P*-value, and conclude in context.

b. Can you conclude that tongue piercing causes more enamel cracks? Explain your reasoning.

18.32 Distraction and snacking. A study assessed the effects of playing a computer game (solitaire) during lunch on behavioral and physiological variables. A sample of 44 healthy adults was randomly assigned to eat lunch either with the computer game distraction or without distraction. The food served for lunch was exactly the same for everyone. All participants also had the same breakfast. Thirty minutes after lunch, participants were offered cookies as a snack and were instructed to eat as many or as few cookies as they liked. Here are the summary statistics for the findings:[25]

Group	**n**n	**x-**$\bar{x}$	**s**s
Distraction	22	52.1	45.1
No distraction	22	27.1	26.4

a. Do the findings support the hypothesis of greater snack intake after meals taken with a distraction? State hypotheses in terms of the two population means, obtain the two-sample *t* statistic and *P*-value, and conclude using a significance level of 0.05.

b. Can you conclude that distraction during meals causes the greater snack intake? Explain your reasoning.

18.33 Umbilical cord clamping. Could preventing iron deficiency in infants by delaying the clamping of the umbilical cord help promote child development? Healthy Swedish newborns were randomly assigned to have their umbilical cord clamped either right after birth or after a 3-minute delay. Four years later, parents were instructed to administer the Ages and Stages Questionnaire (ASQ) to assess psychomotor development in their child. The ASQ is scored on a scale of 0 to 300, with higher scores indicating more developmental milestones reached. Here are the summary findings:[26]

Group	**n**n	**x-**$\bar{x}$	**s**s
Delayed clamping	130	278.9	21.6
Early clamping	115	275.5	27.6

Obtain a 95% confidence interval for the difference in mean ASQ score for the populations of 4-year-old Swedish children with delayed and early cord clamping at birth.

18.34 Umbilical cord clamping, continued. Based on the study described in Exercise 18.33, we want to know if 4-year-old Swedish children with delayed cord clamping are developmentally more advanced, scoring higher on average on the ASQ than 4-year-olds with early cord clamping.

a. Write the corresponding null and alternative hypotheses.
b. Use the confidence interval you computed in Exercise 18.33 to reject or fail to reject the null hypothesis. Explain your answer.

18.35 Acupuncture for treating migraines. Acupuncture is widely used to prevent migraine attacks, but does it work? A study investigated the effectiveness of acupuncture compared with sham acupuncture and with no acupuncture (patients left on the wait list) in individuals with migraines. Patients were randomly assigned to one of the three “treatments” for a period of eight weeks, after which they monitored their migraine attacks over the next three weeks. Here are summary data for the number of days in which each patient used pain medication for migraines or headaches during the monitoring period:[27]

Acupuncture			**Sham**			**Wait list**		
n	$\bar{x}$	s	n	$\bar{x}$	s	n	$\bar{x}$	s
132	3.2	3.0	76	3.4	2.9	64	4.4	3.6

a. Did patients who received the acupuncture treatment have significantly fewer pain medication days than patients who stayed on the wait list? Perform the appropriate hypothesis test and conclude in context.
b. Give a 95% confidence interval for the mean difference. Describe the effect size. Would you consider recommending acupuncture for the preventive treatment of headaches?

18.36 Acupuncture for treating migraines, continued. Go back to Exercise 18.35. The placebo effect has been shown to be particularly strong in the treatment of pain.

a. Is there significant evidence that patients who received the acupuncture treatment had significantly fewer pain medication days than patients who received the sham acupuncture (needles inserted at random locations)? Perform the appropriate hypothesis test and conclude in context.
b. Explain what makes this study a randomized, controlled

experiment. Finding statistical significance in Exercise 18.35 does not imply that acupuncture is causing the improvement. Explain briefly why.

18.37 Acupuncture for treating migraines: critique. The initial pool of volunteers for the study described in Exercise 18.35 was larger. A few individuals dropped out for various reasons after the study started. Here are the retention rates for all three groups:

Acupuncture	Sham	Wait list
91%	94%	84%

Comparatively more patients dropped out of the study in the wait list group. This could be just chance; or maybe individuals with fewer migraines lost patience with being left on the wait list for several weeks; or maybe those who least believed in the effectiveness of acupuncture decided to drop out. How could this fact affect the conclusions of the study?

The following exercises ask you to analyze data and conclude in context without explicit guidance. Follow the ***Plan, Solve,*** *and* ***Conclude*** *steps of the four-step process. It may be helpful to restate in your own words the* ***State*** *information given in the exercise.*

18.38 Student drinking. A professor asked her sophomore students, "How many drinks do you typically have per session?" (A drink is defined as one 12-ounce beer, one 4-ounce glass of wine, or one 1-ounce shot of liquor.) Some of the students didn't drink. Table 18.4 gives the responses of the female and male students who did drink.[28] It is likely that some of the students exaggerated a bit. The sample is all students in one large sophomore-level class. The class is popular, so we are tentatively willing to regard its members as an SRS of sophomore students at this college. Do a complete analysis that reports on

a. the drinking behavior claimed by sophomore women.

b. the drinking behavior claimed by sophomore men.

c. a comparison of the behavior of women and men.

TABLE 18.4 **Drinks per session claimed by female and male students**

Female students							
2.5	9	1	3.5	2.5	3	1	3
5	3.5	5	1	2	1	7	3
3	6	5	3	8	6	6	3
4	5	3.5	4	2	1	5	5
7	7	7	5.5	3	2.5	10	5
2	5	2.5	3	4.5	9	5	4
4	5	1	5	3	4	10	7
2	1	2.5	2.5				

Male students							
7	7.5	8	15	3	4	1	5
16	4	8	5	9	7	7	3
6	8	8	4.5	10.5	8	6	10
15	3	10	7	4	6	5	2
7	3	7	6	4	5	2	5
8	4	2	4	12.5	3	15	2
6	4.5	5					

18.39 Ink toxicity. The National Toxicology Program evaluates the toxicity of chemicals found in manufacturing, in consumer products, or in the environment after disposal. Toxicity is assessed through a battery of tests. Here are some results from a study of the toxicity of black newsprint ink in 7-week-old female rats. The rats’ fur was locally clipped twice a week for 13 weeks. One group of rats received a dermal application of ink right after each clipping, and a control group of rats was left untreated. Table 18.5 shows the body weights (in grams) of female rats at the beginning of the study and at the end of the 13 weeks.[29]

a. Verify that the two experimental groups are not significantly different at the beginning of the study.

b. Is there good evidence that ink application impairs growth in female rats between 7 and 20 weeks of age? Estimate the difference between the mean weight gains of the two populations. (Use a 95% confidence level.)

TABLE 18.5 Body weights (g) at beginning and end of ink toxicity study

Control Group		Treatment Group	
Week 0	**Week 13**	**Week 0**	**Week 13**
111.2	191.6	107.3	187.0
105.4	191.2	116.7	189.5
110.8	210.7	112.2	179.2
105.6	185.2	103.4	172.2
106.1	195.0	113.2	178.7
104.4	188.3	110.6	180.9
114.0	188.4	110.6	188.3
115.1	195.6	100.5	188.9
109.2	204.6	106.3	183.1
111.3	195.7	112.5	184.5

18.40 Magnets for pain relief. A randomized, double-blind experiment studied whether magnetic fields applied over a painful area can reduce pain intensity. The subjects were 50 volunteers with postpolio syndrome who reported muscular or arthritic pain. The pain level when pressing a painful area was graded subjectively on a scale from 0 to 10 (where 0 is no pain and 10 is maximum pain).

Patients were randomly assigned to wear either a magnetic device or a placebo device over the painful area for 45 minutes. Here is a summary of the pain scores for this experiment, expressed as means ±± standard deviations:[30]

	Magnetic device (n == 29)	Placebo device (n == 21)
Pretreatment	9.6 ±± 0.7	9.5 ±± 0.8
Post-treatment	4.4 ±± 3.1	8.4 ±± 1.8
Change	5.2 ±± 3.2	1.1 ±± 1.6

a. Is there good evidence that the magnetic device is better than a placebo device at reducing pain?

b. How much reduction in pain is achieved with the magnetic device? Give a 95% confidence interval for the mean difference in pain scores pre- and post-treatment among patients given the magnetic device. What procedure did you use and why?

18.41 Magnets for pain relief, continued. Conclusions from any study should be more comprehensive than a simple statement of significance or a confidence interval.

a. The random assignment of subjects to treatments can sometimes lead, by chance, to an unbalanced split of subjects. Is there evidence of a significant difference in mean pain scores between the two groups at the beginning of the experiment?

b. You have used summary data from this experiment for your inference. What would you like to know about the raw data to support the legitimacy of your statistical results?

18.42 The cost of sex. Evolution theory predicts a physiological cost of increased reproduction that translates into a shorter life span. A study examined the cost of sexual activity in male fruit flies by comparing males allowed to be sexually active with those that were not. Table 18.6 gives the longevity in days for a sample of males randomly assigned to one or the other condition.[31] Do a complete analysis that reports on

a. the longevity of sexually active male fruit flies.

b. the longevity of sexually inactive male fruit flies.

c. a comparison of the longevity of sexually active and sexually inactive male fruit flies.

TABLE 18.6 **Longevity (days) of male *Drosophila***

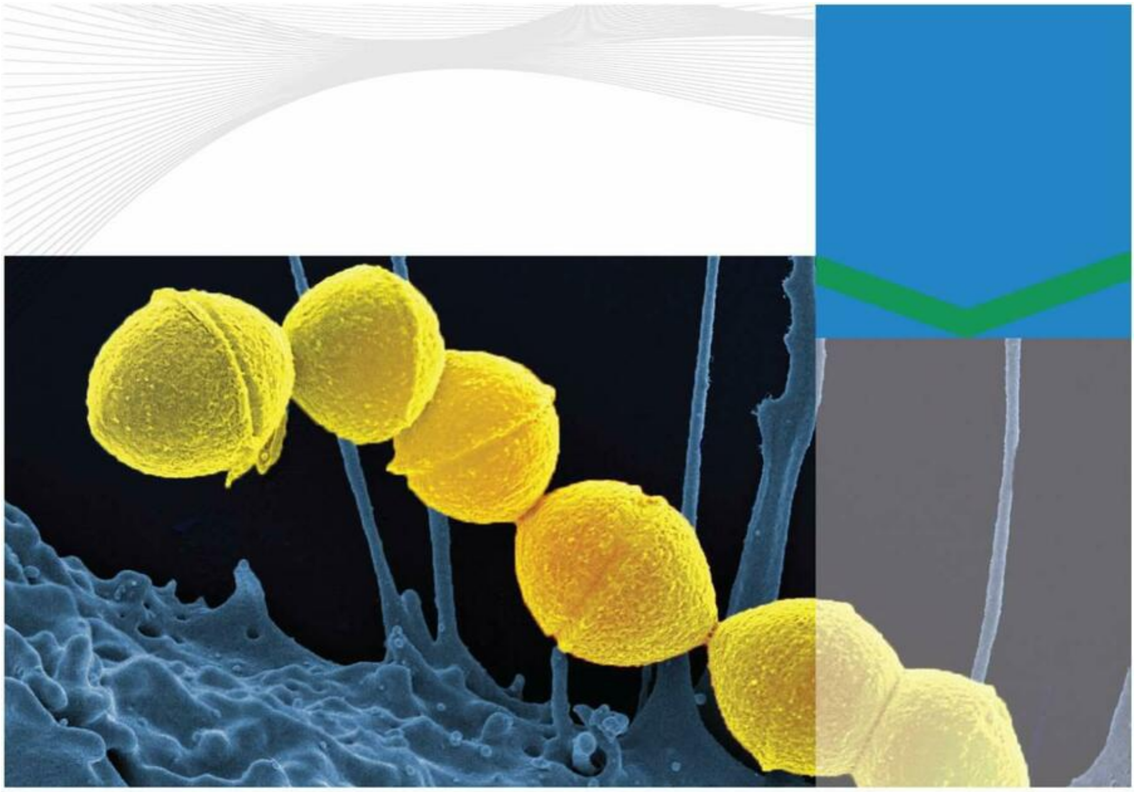
Sexually inactive					Sexually active				
35	49	64	70	76	16	30	35	42	54
37	56	65	70	76	19	33	35	44	54
39	56	65	70	81	19	34	40	46	54
46	64	65	76	85	26	34	42	46	56

46	64	70	76		30	34	42	46	61

Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book’s review chapters (Chapters 8, 16, and 25). Note that short answers are not available for these exercises to enable their use for class work or assignments.*

18.43 Elderly health. Go back to Exercise 17.47 (page 444). Do elderly men and women differ, on average, in their blood calcium levels? Do they differ in their blood inorganic phosphorus levels? Do a complete analysis.

CHAPTER 19 Inference about a Population Proportion



National Institute of Allergy and Infectious Diseases, National Institutes of Health

IN THIS CHAPTER WE COVER...

- The sample proportion p^$\hat{p}$
- Large-sample confidence intervals for a proportion
- Accurate confidence intervals for a proportion
- Choosing the sample size
- Hypothesis tests for a proportion

Our discussion of statistical inference up to this point has concerned making inferences about population *means* when working with quantitative data. Now we turn to questions about the *proportion* of some outcome in a population when studying a categorical variable. Here are some examples that call for inference about population proportions.

EXAMPLE 19.1 Assessing the prevalence of HPV infections

Human papillomavirus (HPV) infection is the most common sexually transmitted infection. Certain types of HPV can cause genital warts in both men and women and cervical cancer in women. Whether or not to vaccinate all young women against HPV before they become sexually active is still the subject of heated debate. But just how common is HPV infection in women? The U.S. National Health and Nutrition Examination Survey (NHANES) contacted a representative sample of 1921 women between the ages of 14 and 59 years and asked them to provide a self-collected vaginal swab specimen. Of these 1921 specimens, 515 tested positive for HPV, indicating a current HPV infection. That’s 26.8% of the sample.[1] Based on these data, what can we say about the percent of all women aged 14 to 59 who test positive for HPV? We want to *estimate a single population proportion.* This chapter concerns inference about one proportion.

EXAMPLE 19.2 How to quit smoking

If you want to quit smoking, is it better to stop abruptly or gradually? Researchers in England recruited adults who smoked at least 15 cigarettes daily and were willing to quit smoking. The subjects were randomly assigned to either the abrupt-cessation group or the gradual-cessation group. Both groups received similar support in the form of behavioral counseling and nicotine patches. Verified abstinence was achieved four weeks later by 39.2% of the gradual-cessation group and 49.0% of the abrupt-cessation group.[2] Is this significant evidence that the proportions of smokers who still abstain from smoking four weeks after they stopped smoking differ in the population of all smokers who quit gradually and the population of all smokers who quit abruptly? We want to *compare two population proportions.* This is the topic of Chapter 20.

STATS IN YOUR WORLD

WHO IS A SMOKER?

When estimating a proportion p,p, make sure you know what counts as a “success.” The news says that 20% of adolescents smoke—a shocking proportion. It turns out that this is the percent who smoked at least once in the past month. If we say that a smoker is someone who smoked on at least 20 of the past 30 days and smoked at least half a pack on those days, fewer than 4% of adolescents qualify.

To do inference about a population mean μ,μ, we used the mean x-$\bar{x}$ of a random sample from the population. The reasoning of inference started with the sampling distribution of x-.$\bar{x}$. Now we follow the same pattern, replacing means with proportions.

THE SAMPLE PROPORTION p^$\hat{p}$

We are interested in the unknown proportion pp of a population that has some outcome. For convenience, call the outcome we are looking for a “success.” In Example 19.1 the population is women aged 14 to 59, and the parameter pp is the proportion who currently test positive for HPV. To estimate p,p, NHANES used a nationally representative sample of 1921 women. In this sample, 515 women tested positive for HPV. The statistic that estimates the parameter pp is the **sample proportion:**

p^=number of successes in the sampletotal number of individuals in the sample=5151921=0.2681

$$\begin{aligned}\hat{p} &= \frac{\text{number of successes in the sample}}{\text{total number of individuals in the sample}}\\ &= \frac{515}{1921} = 0.2681\end{aligned}$$

Read the sample proportion p^$\hat{p}$ as “p-hat.”

How good is the statistic p^$\hat{p}$ as an estimate of the parameter p?p? To find out, we ask, “What would happen if we took many random samples?” We first explored this idea in Example 13.9 (page 337) and the corresponding Figure 13.8. The sampling distribution of p^$\hat{p}$ formalizes this question.

We described the properties of the sampling distribution of p^$\hat{p}$ in Chapter 13 (page 338), and Figure 13.9 (page 339) displays them graphically. The behavior of sample proportions p^$\hat{p}$ is similar to the behavior of sample means x-.$\bar{x}$. When the sample size nn is large, the sampling distribution is approximately Normal. The larger the sample, the more nearly Normal the distribution is. *Don’t use the Normal approximation to the distribution of* p^$\hat{p}$ *when the sample size* nn *is small.*

CAUTION

The mean of the sampling distribution of p^$\hat{p}$ is the true value of the population proportion p.p. That is, p^$\hat{p}$ is an unbiased estimator of p.p. The standard deviation of p^$\hat{p}$ is p(1-p)/n,$\sqrt{p(1-p)/n}$, which becomes smaller as the sample size nn is increased, so that estimation is likely to be more accurate when the sample is larger. As is the case for x-,$\bar{x}$, the standard deviation gets smaller only at the rate n.$\sqrt{n}$. We need four times as many observations to cut the standard deviation in half, and 100 times as many observations to cut it 10 times.

The Normal approximation to the sampling distribution of p^$\hat{p}$ *is least accurate when* pp *is close to 0 or 1*. If p=0,$p = 0$, successes are impossible; every sample has p^=0,$\hat{p} = 0$, and no Normal distribution can model that. In the same way, the approximation works poorly when pp is close to 1. In practice, we need much larger nn for values of pp near 0 or 1.

Inference about a population proportion pp starts with using the sample proportion p^$\hat{p}$ to estimate p.p. Confidence levels and *P*-values are probabilities calculated from the sampling distribution of p^.$\hat{p}$. We will consider only situations that allow us to use the Normal approximation to this sampling distribution. Here is a summary of the conditions we need.

CONDITIONS FOR INFERENCE ABOUT A PROPORTION

- We can regard our data as a **simple random sample** (SRS) from the population. As usual, this is the most important condition.
- The **sample size** nn **is large enough** to ensure that the sampling distribution of p^$\hat{p}$ is close to Normal. We will see that different inference procedures require different answers to the question "How large is large enough?"

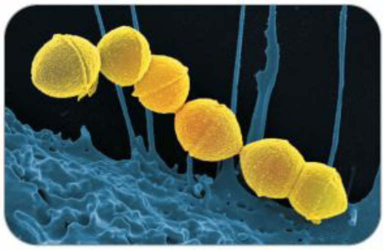
All the inference procedures covered in this book also require that the population be much larger than the sample.[3] This condition is usually satisfied in practice and is satisfied in all our examples and exercises.

APPLY YOUR KNOWLEDGE

19.1 Antibiotic resistance. A sample of 1714 cultures from individuals in Florida diagnosed with a strep infection was tested for resistance to the antibiotic penicillin; 973 cultures showed partial or complete resistance to the antibiotic.[4]

a. Describe the population and explain in words what the parameter pp is.

b. Give the numerical value of the statistic p^$\hat{p}$ that estimates p.p.



National Institute of Allergy and Infectious Diseases, National Institutes of Health

19.2 Coffee drinkers. A July 2015 Gallup poll interviewed by phone a random sample of 1009 American adults; 675 said they were regular coffee drinkers.

a. Describe the population and explain in words what the parameter pp is.

b. Give the numerical value of the statistic p^$\hat{p}$ that estimates p.p.

19.3 No inference. An article published on National Public Radio’s website in 2012 discussed the recent advances in genome sequencing and offered an opt-in poll. The question asked was “Would you have your genome sequenced if you could afford it?” The poll stayed open for two weeks, during which time 6627 clicks were received (5398 for yes, 656 for undecided, and 573 for no).[5] We can’t use these data as the basis for inference about the proportion of all Americans who would have their genomes sequenced if they could afford it. Why not? Could we use these data as the basis for inference about the proportion of all NPR online readers who would have their genomes sequenced if they could afford it?

LARGE-SAMPLE CONFIDENCE INTERVALS FOR A PROPORTION

To estimate a population proportion p,p, use the sample proportion p^.$\hat{p}$. If the conditions for inference apply, the sampling distribution of p^$\hat{p}$ is close to Normal, with mean pp and standard deviation p(1-p)/n.$\sqrt{p(1-p)/n}$.

We saw in Chapter 14 (page 354) that a level CC confidence interval for a parameter, based on the sampling distribution of a statistic estimate, takes the form “estimate ±$\pm$ margin of error,” where the margin of error is computed as a critical value times the standard deviation of the sampling distribution of the estimate. To obtain a level CC confidence interval for p, p, we would therefore like to use the formula

p^±z*p(1-p)n

$$\hat{p} \pm z^* \sqrt{\frac{p(1-p)}{n}}$$

with the critical value z*z^* chosen to cover the central area CC under the standard Normal curve. Figure 19.1 shows why.

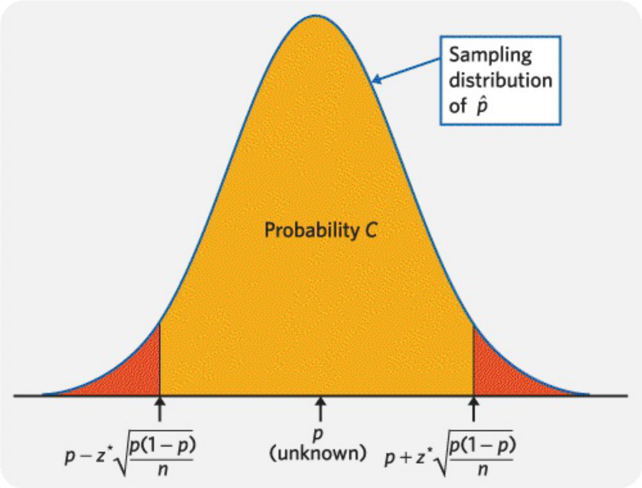
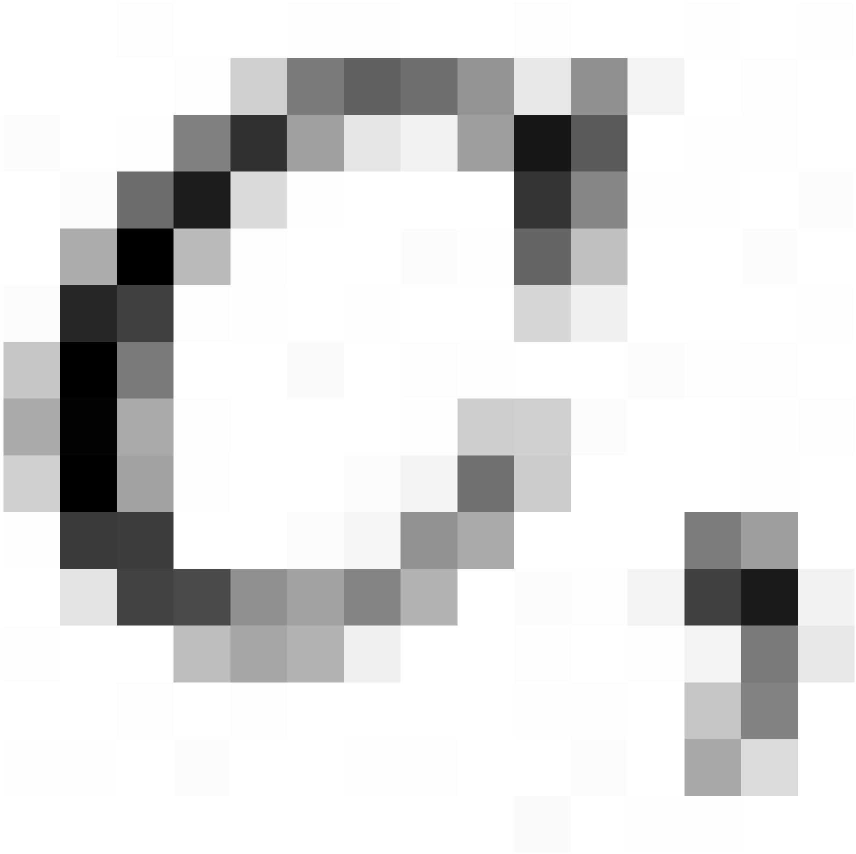
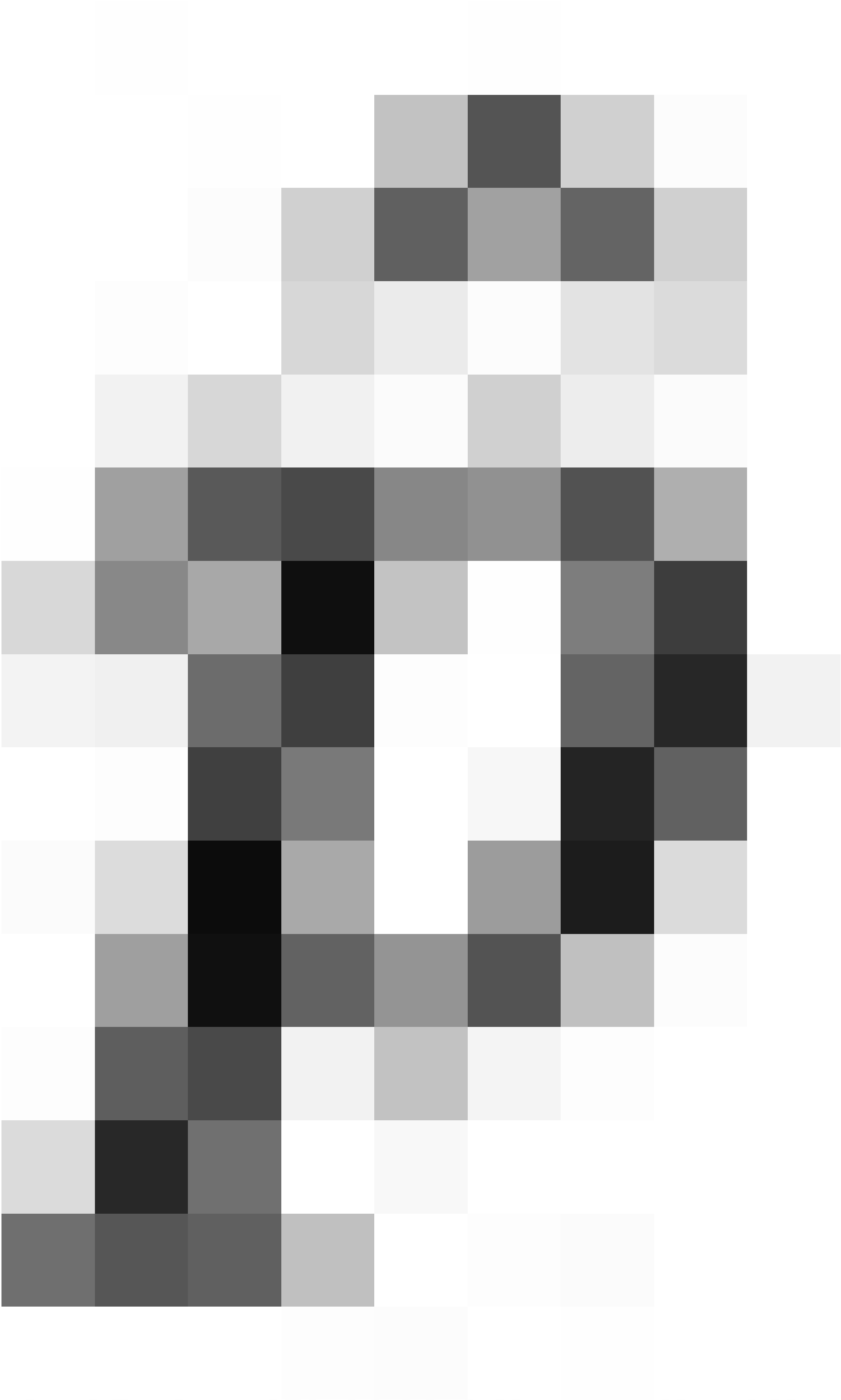
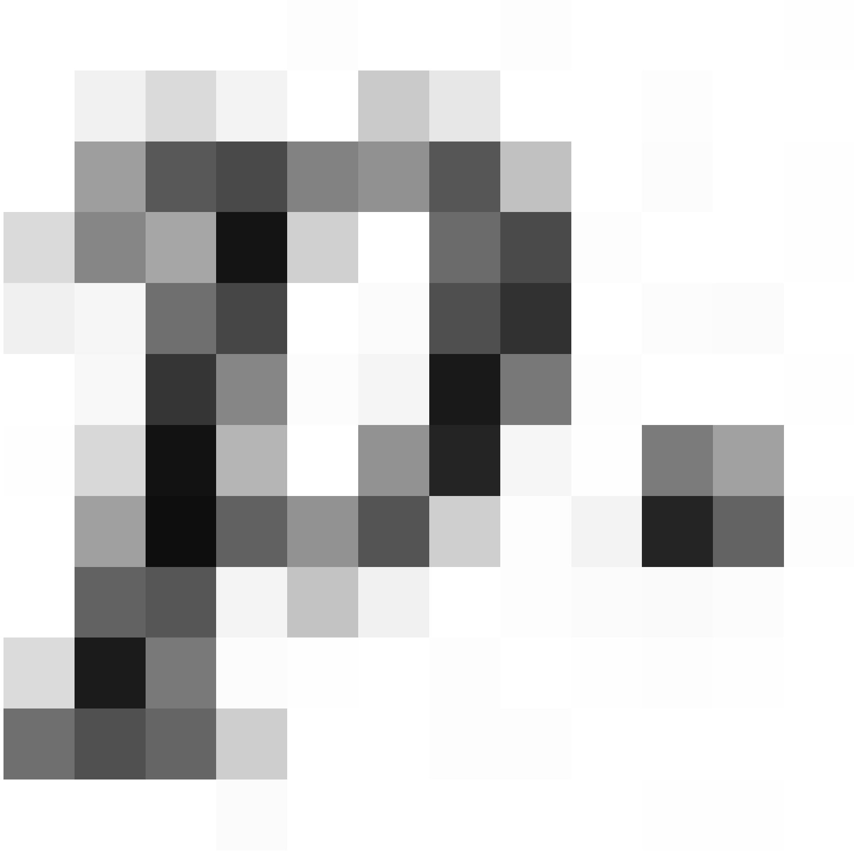


Figure 19.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 19.1 With probability C,

 the random variable

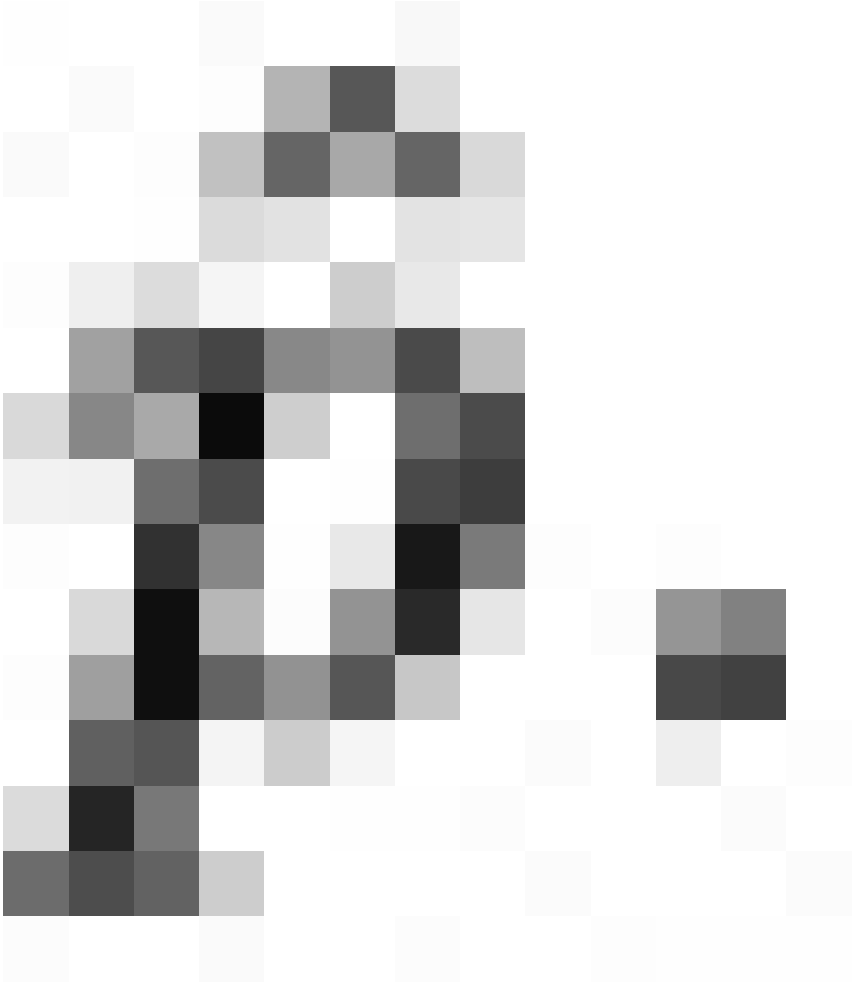
p^  lies within ±z*p(1-p)/n $\pm z^* \sqrt{p\left(1-p\right)/n}$ of the unknown population proportion p.



In other words, in

these samples p p lies

within ±z*p(1-p)/n $\pm z^* \sqrt{p\left(1-p\right)/n}$ of

p^.


The major challenge of constructing a confidence interval for a population proportion is that we don’t know the value of p.p. One approach replaces the standard deviation of p^$\hat{p}$ with the p^$\hat{p}$ **standard error of** p^$\hat{p}$

SEp^=p^(1-p^)n

$$\mathrm{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

to get the confidence interval

p^±z*p^(1-p^)n

$$\hat{p} \pm z^{*}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This interval has the form

estimate±z*SEestimate

$$\text{estimate} \pm z^{*}\,\mathrm{SE}_{\text{estimate}}$$

We can trust such a confidence interval only for large samples. Because the number of successes must be a whole number, using a continuous Normal distribution to describe the behavior of p^$\hat{p}$ may not be accurate unless nn is large. Because the approximation is least accurate for populations that are almost all successes or almost all failures, we require that the sample have both enough successes and enough failures rather than that the overall sample size be large. *Pay attention to both conditions for inference in the box that summarizes the confidence interval: We must be willing to regard the sample as an SRS from the population, and the sample must have both enough successes and enough failures.*

CAUTION

LARGE-SAMPLE CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

Draw an SRS of size nn from a large population that contains an unknown proportion pp of successes. An approximate level CC **confidence interval for** pp is

p^±z*p^(1-p^)n

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where z*z^* is the critical value for the standard Normal density curve with area CC between -z*$-z^*$ and z*.z^*. Technology with basic statistical inference capabilities will perform all necessary computations and produce the lower and upper boundaries of the confidence interval.

Use this interval only when the number of successes and the number of failures in the sample are each at least 15.[6]

Why not *t*? Unlike in Chapter 17, here we *don't* change z*z^* to t*t^* when we replace the standard deviation of the estimate by its standard error. When the sample mean x-$\bar{x}$ estimates the population mean μ,μ, a separate parameter σσ describes the spread of the distribution of x-.$\bar{x}$. We separately estimate σ,σ, and this can be shown mathematically to lead to a *t* distribution. However, when the sample proportion p^$\hat{p}$ estimates the population proportion p,p, the spread of the sampling distribution depends on p,p, not on a separate parameter. There is therefore no *t* distribution—we just make the Normal approximation a bit less accurate when we replace pp in the standard deviation by p^.$\hat{p}$.

The next section introduces a more accurate method for computing a confidence interval for a population proportion, and we recommend that you always use this more accurate method. Nevertheless, you may occasionally come across publications that use the large-sample method, so here is an example to help you understand how it works.

EXAMPLE 19.3 Assessing the prevalence of HPV infections

The four-step process for obtaining a confidence interval is outlined on page 354.

STATE: NHANES found that 515 of a sample of 1921 women aged 14 to 59 years currently tested positive for HPV. That is, p^=0.2681. $\hat{p} = 0.2681$. What can we say about HPV prevalence in the population of all women aged 14 to 59 in the United States?

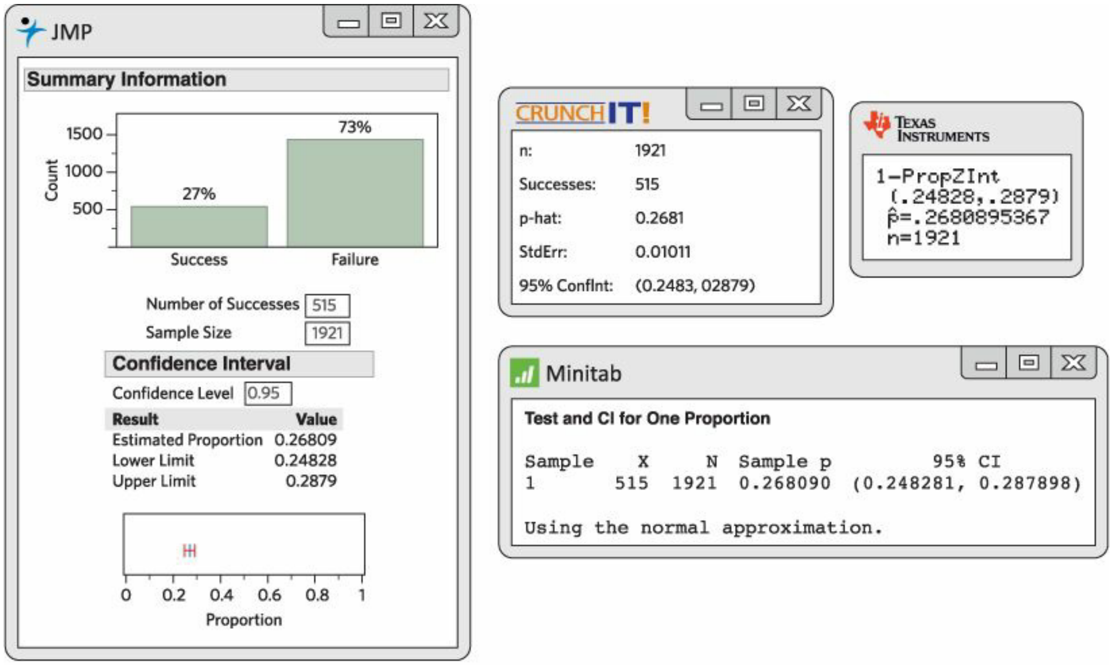
PLAN: We will give a 95% confidence interval to estimate the proportion pp of all women aged 14 to 59 in the United States who currently test positive for HPV.

SOLVE: First verify the conditions for inference:

- The sampling design was a complex stratified sample, and the survey used inference procedures for that design. The overall effect is close to an SRS, however.
- The sample is large enough: The numbers of successes (515) and failures (1406) in the sample are both much larger than 15.

The sample size condition is easily satisfied. The condition that the sample be an SRS is only approximately met.

We now use technology to obtain a confidence interval for p.p. Figure 19.2 gives (rounded to various decimal places) the sample estimate p^=0.2681$\hat{p} = 0.2681$ and the lower and upper bounds of a 95% confidence interval: (0.2483, 0.2879).


JMP
Summary Information
Count
1500
1000
500
27%
73%
Success
Failure
0
0.2
0.4
0.6
0.8
1
Proportion
CRUNCH IT!
TEXAS INSTRUMENTS
Minitab

```
Number of Successes  515
Sample Size          1921
Confidence Interval
Confidence Level  0.95
Result                Value
Estimated Proportion  0.26809
Lower Limit           0.24828
Upper Limit           0.2879
```

```
n:            1921
Successes:    515
p-hat:        0.2681
StdErr:       0.01011
95% ConfInt:  (0.2483, 02879)
```

```
1-PropZInt
 (.24828,.2879)
 p̂=.2680895367
 n=1921
```

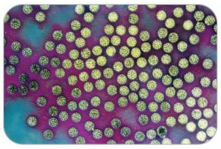
```
Test and CI for One Proportion

Sample    X     N  Sample p          95% CI
1       515  1921  0.268090  (0.248281, 0.287898)

Using the normal approximation.
```

Figure 19.2

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 19.2 Output from a graphing calculator and three statistical software programs for the 95% confidence interval for the proportion of women ages 14 to 59 who test positive for HPV.

CONCLUDE: We are 95% confident that the percent of women aged 14 to 59 in the United States who are currently experiencing an HPV infection (the percent who would receive a positive result if tested) lies somewhere between 24.8% and 28.8%.



Kwangshin Kin/Science Source

The computations performed by software in this example involve the following steps:

- Sample estimate: p^=515/1921=0.2681$\hat{p} = 515/1921 = 0.2681$
- Standard error of p^:$\hat{p}$:

SE=p^(1-p^)n=(0.2681)(0.7319)1921=0.01011

$$\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.2681)(0.7319)}{1921}} = 0.01011$$

- Standard Normal critical value for a 95% confidence level: z*=1.960 $z^* = 1.960$ (rounded to three decimal places, obtained from software or Table C)
- Margin of error: m=z*$m = z^*$ SE=1.960(0.01011)=0.0198 $\text{SE} = 1.960\,(0.01011) = 0.0198$
- 95% confidence interval for p:p^±m=0.2681±0.0198, $p : \hat{p} \pm m = 0.2681 \pm 0.0198$, or 0.2483 to 0.2879

Not all software packages offer the zz procedures for a proportion. Excel and SPSS, for instance, do not have an easy way to perform these procedures, and R uses a slightly different methodology. The graphing calculator and three statistical programs in Figure 19.2 all use the Normal approximation for the sampling distribution of p^,$\hat{p}$, although this is something you must specify with Minitab.

Interpreting a confidence interval in context presents the same challenges whether the data are quantitative or categorical, and it involves much more than simple formulas for inference. For instance, the practical problems of a large sample survey somewhat weaken our trust in the NHANES conclusions. Only women who agreed to provide a self-collected vaginal swab and collected it properly were included in the final sample. According to the study authors, nonrespondents represented 23% of the original sample and tended to be younger and members of ethnic minority groups. A nonresponse rate of 23% is not unusual in large sample surveys, but the fact that nonrespondents were somewhat different from respondents overall may cause some bias in the sample results. To address this issue, the survey used statistical methods that adjust for unequal response rates in different groups.[7]

APPLY YOUR KNOWLEDGE

19.4 Coffee drinkers, continued. Exercise 19.2 presented data about people who drink coffee regularly. Give a 95% confidence interval for the proportion of American adults who are regular coffee drinkers. Follow the four-step process as illustrated in Example 19.3.

19.5 Antibiotic resistance, continued. Exercise 19.1 gave data about the resistance of strep cultures to the antibiotic penicillin. Give a 95% confidence interval for the proportion of strep cultures from Florida patients showing partial or complete resistance to penicillin. Follow the four-step process as illustrated in Example 19.3.

19.6 Who is asking? A Harris Interactive Poll asked the same questions to two samples of adults: One group answered the survey on the phone with an interviewer and the other completed the survey online by themselves. When asked if they exercised regularly, 58% of those interviewed on the phone said “yes,” compared with 35% of those who completed the survey online.

a. What kind of answer variability or bias is most likely to produce such different results?

b. Computing a 95% confidence interval for the true proportion of adult Americans who exercise regularly gives very different results depending on which sample is used. Is this kind of variability what the margin of error quantifies? Briefly explain your answer.

19.7 Clinical progression of SARS. Severe acute respiratory syndrome (SARS) is a viral respiratory illness that was first reported in Asia in 2003 and then spread worldwide to near pandemic levels before eventually being contained. A study examined a random sample of 75 patients diagnosed with SARS to describe the disease’s clinical progression. One finding was that, while fever and pneumonia initially improved after treatment in all 75 patients, 64 patients developed recurrent fever.[8] Explain why we can’t use the large-sample confidence interval to estimate the proportion pp of the population of patients treated for SARS who develop recurrent fever.

Rocky Mountain Laboratories,
National Institute of Allergy and
Infectious Diseases, NIH

ACCURATE CONFIDENCE INTERVALS FOR A PROPORTION

The confidence interval p^±z*p^(1-p^)/n$\hat{p} \pm z^* \sqrt{\hat{p}\,(1-\hat{p})\,/n}$ for a population proportion pp is easy to calculate. It is also easy to understand, because it rests directly on the approximately Normal distribution of p^.$\hat{p}$. Unfortunately, confidence levels from this interval are often quite inaccurate unless the sample is very large. Simulations show that the actual confidence level is usually *less* than the confidence level you asked for in choosing the critical value z*.z^*. That's bad. What is worse, accuracy does not consistently get better as the sample size nn increases. Both "lucky" and "unlucky" combinations of the sample size nn and the true population proportion pp may occur.

Fortunately, a simple modification has been shown experimentally to successfully improve the accuracy of the confidence interval. We call it the "plus four" method, because all you need to do is *add four imaginary observations, two successes and two failures.* With the added observations, the **plus four estimate** of pp is

p~=number of successes in the sample+2n+4

$$\tilde{p} = \frac{\text{number of successes in the sample} + 2}{n+4}$$

The formula for the confidence interval is exactly as before, with the new sample size and the new number of successes.[9] You do not need software that offers the plus four interval—just enter the new sample size (actual size + 4) and the new number of successes (actual number + 2) into the large-sample procedure.

PLUS FOUR CONFIDENCE INTERVAL FOR A PROPORTION

Draw an SRS of size nn from a large population that contains an unknown proportion pp of successes. To get the **plus four confidence interval for** pp**,** add four imaginary observations, two successes and two failures. Then use the large-sample confidence interval with the new sample size (n+4)$(n+4)$ and the new count of successes (actual count $+$ 2).

Use this interval when the confidence level is at least 90% and the sample size n is at least 10, with any counts of success and failure.

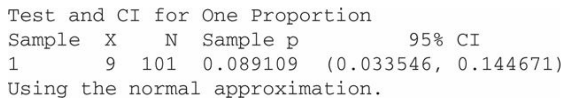
EXAMPLE 19.4 Blinding in clinical trials

STATE: Many clinical trials randomly assign patients to either an active treatment or a placebo. These trials are always double-blind. Sometimes, however, the patients may be able tell whether they are getting the active treatment, which defeats the purpose of blinding. Reports of medical research usually ignore this problem. Investigators looked at a random sample of 97 articles reporting on placebo-controlled randomized trials in the top five general medical journals. Only 7 of the 97 discussed the success of blinding—and in 5 of these studies, the blinding was imperfect.[10] What proportion of all such studies discuss the success of blinding?

PLAN: Take p to be the proportion of articles that discuss the success of blinding. Give a 95% confidence interval for p.

SOLVE: We are told that the researchers selected a random sample of articles, so the condition that the data be an SRS appears to be met. The condition for use of the large-sample interval is not met, however, because there are fewer than 15 successes in the sample. Fortunately, the total sample size $(n = 97)$ is large enough for a plus four confidence interval.

Here is the Minitab output when we add two successes and two failures to the original data, so that the new number of successes is $7 + 2 = 9$ and the new sample size is $97 + 4 = 101$.



```
Test and CI for One Proportion
Sample  X    N  Sample p          95% CI
1       9  101  0.089109  (0.033546, 0.144671)
Using the normal approximation.
```

This plus four confidence interval, (0.0335, 0.1447), is the same as a large-sample interval that would be based on 9 successes in 101 observations.

> **CONCLUDE:** We estimate with 95% confidence that between 3.4% and 14.5% of all articles reporting on placebo-controlled clinical trials discuss whether the blinding succeeded.

Here are the computations performed by Minitab for this example:

- Plus four estimate of p:p :

  p̃=7+297+4=9101=0.0891

  $$\tilde{p} = \frac{7+2}{97+4} = \frac{9}{101} = 0.0891$$

- Standard error of p̃:$\tilde{p}$:

  SE=p̃(1-p̃)n+4=(0.0891)(0.9109)101=0.02835

  $$\mathrm{SE} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \sqrt{\frac{(0.0891)(0.9109)}{101}} = 0.02835$$

- Standard Normal critical value for a 95% confidence level: z*=1.960 $z^* = 1.960$ (rounded to three decimal places, obtained from software or Table C)
- Margin of error: m=z*$m = z^*$ SE=1.960(0.02835)=0.05556 $\mathrm{SE} = 1.960\,(0.02835) = 0.05556$
- Plus four 95% confidence interval for p:p̃±m=0.0891±0.0556, $p : \tilde{p} \pm m = 0.0891 \pm 0.0556,$ or 0.0335 to 0.1447

For comparison, the ordinary sample proportion in Example 19.4 is p^=7/97=0.0722.$\hat{p} = 7/97 = 0.0722.$ The plus four estimate p̃=0.0891$\tilde{p} = 0.0891$ is farther away from zero than p^=0.0722.$\hat{p} = 0.0722.$ The plus four estimate gains its added accuracy by always moving toward 0.5 and away from 0 or 1, whichever is closer. This is particularly helpful when the sample contains only a few successes or a few failures. The numerical difference between a large-sample interval and the corresponding plus four interval is often small. Remember that the confidence level is the probability that the interval will catch the true population proportion *in very many uses*. Small differences every time add up to accurate confidence levels from plus four intervals versus inaccurate levels from large-sample intervals.

How much more accurate is the plus four interval? Computer studies have asked how large nn must be to guarantee an accurate confidence interval. If the true population proportion is p=0.1,$p = 0.1,$ for example, the answer is n≥646$n \geq 646$ for the large-sample interval and n≥11$n \geq 11$ for the plus four interval when we want a 95% confidence level.[11] The consensus of computational and theoretical studies is that plus four is very much better than the large-sample interval for many combinations of nn and p.$p.$ **We recommend that you always use the plus four interval for estimating a proportion.**

APPLY YOUR KNOWLEDGE

19.8 Clinical progression of SARS, continued. Go back to Exercise 19.7, which describes the progression of fever symptoms in individuals treated for SARS.

a. The plus four method adds four observations, two successes and two failures. What are the sample size and the count of successes after you do this? What is the plus four estimate p$\tilde{p}$ of the proportion pp in the population of patients treated for SARS who develop recurrent fever?

b. Give the plus four 95% confidence interval for p.$p.$

19.9 Antibiotic resistance, continued. In Exercise 19.5 you gave a 95% confidence interval for the proportion of strep cultures from Florida patients showing partial or complete resistance to penicillin. This estimate was based on the finding of 973 cultures with partial or complete resistance out of a random sample of 1714 throat cultures.

a. What is the plus four estimate p$\tilde{p}$ for this population proportion? Give the plus four 95% confidence interval for p.$p.$

b. How do the two confidence intervals compare? Explain why the two methods give similar results when the sample size is large. For large samples, both methods give similar and accurate estimates. For smaller samples, the plus four estimate is always more accurate.

CHOOSING THE SAMPLE SIZE[i]

In planning a study, we may want to choose a sample size that will allow us to estimate the parameter within a given margin of error. We saw earlier (page 391) how to do this for a population mean. The method is similar for estimating a population proportion.

STATS IN YOUR WORLD

KIDS ON BIKES

In the most recent year for which data are available, 77% of children killed in bicycle accidents were boys. You might take these data as a sample and start from p^=0.77$\hat{p}=0.77$ to do inference about bicycle deaths in the near future. What you should not do is conclude from these data that boys on bikes are in greater danger than girls. We don't know how many boys and girls ride bikes —it may be that most fatalities are boys because most riders are boys.

The margin of error in the large-sample confidence interval for pp is

m=z*p^(1-p^)n

$$m=z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Here z*z^* is the standard Normal critical value for the level of confidence we want. Because the margin of error involves the sample proportion of successes p^,$\hat{p}$, we need to guess this value when choosing n.n. Call our guess p*.p^*. Here are two ways to get p*:p^* :

1. Use a guess p*p^* based on a pilot study or on past experience with similar studies. You can perform several calculations to cover the range of values of p^$\hat{p}$ you might get.
2. Use p*=0.5$p^*=0.5$ as the guess. The margin of error mm is largest when p^=0.5,$\hat{p}=0.5$, so this guess is conservative in the sense that if we get any other p^$\hat{p}$ when we do our study, the margin of error will be smaller than planned.

Once you have a guess p*,p^*, the expression for the margin of error can be solved to give the sample size nn needed. The result for the large-sample confidence interval appears in the following box. For simplicity, use this result even if you plan to use the plus four interval.

SAMPLE SIZE FOR DESIRED MARGIN OF ERROR

The level CC confidence interval for a population proportion p p will have a margin of error approximately equal to a specified value mm when the sample size is

n=z*m2p*(1-p*)

$$n = \left(\frac{z^*}{m}\right)^2 p^* \left(1 - p^*\right)$$

where p*p^* is a guessed value for the sample proportion. The margin of error will be less than or equal to mm if you take the guess p*p^* to be 0.5.

Which method for finding the guess p*p^* should you use? The value of nn you obtain doesn't change much when you change p*p^* as long as p*p^* is not too far from 0.5. You can use the conservative guess p*=0.5$p^* = 0.5$ if you expect the true p^$\hat{p}$ to be roughly between 0.3 and 0.7. If the true p^$\hat{p}$ is close to 0 or 1, using p*=0.5$p^* = 0.5$ as your guess will yield a sample much larger than you need. Try to use a better guess from a pilot study when you suspect that p^$\hat{p}$ will be less than 0.3 or greater than 0.7.

EXAMPLE 19.5 Planning a survey

STATE: Flossing is thought to help prevent tooth decay and gum disease, yet reports suggest that millions of Americans never floss. You are planning a sample survey to determine what percent of students at your university never floss. You obtain permission to use the university roster to contact an SRS of students. You want to estimate the proportion pp of students who never floss with 95% confidence and a margin of error no greater than 4% (0.04). How large a sample do you need?

PLAN: Find the sample size nn needed for margin of error m=0.04 $m = 0.04$ and 95% confidence. Based on your readings on the topic, you guess that the proportion of individuals who never floss is between 30% and 70% for any given population. You can use the guess p*=0.5.$p^* = 0.5$.

SOLVE: The sample size you need is

n=1.960.042(0.5)(1-0.5)=600.25

$$n = \left(\frac{1.96}{0.04}\right)^2 (0.5)(1-0.5) = 600.25$$

Round the result up to n=601.$n = 601$. (Rounding down would yield a margin of error slightly greater than 0.04.)

CONCLUDE: An SRS of 601 students from your university is adequate for a margin of error of ±4%$\pm 4\%$ with a confidence level of 95%.

What sample size would you have needed for a margin of error of 3% rather than 4%? The answer is, after rounding up,

n=1.960.032(0.5)(1-0.5)=1067.11≈1068

$$n = \left(\frac{1.96}{0.03}\right)^2 (0.5)(1-0.5) = 1067.11 \approx 1068$$

For a 2% margin of error, the sample size you would have needed is

n=1.960.022(0.5)(1-0.5)=2401

$$n = \left(\frac{1.96}{0.02}\right)^2 (0.5)(1-0.5) = 2401$$

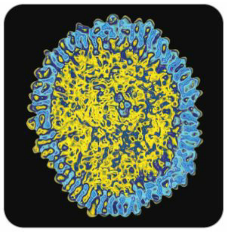
As usual, smaller margins of error call for larger samples. Remember, though, that the population of interest must be much larger than the sample to perform legitimate inference for proportions. Are there more than 24,000 students enrolled at your university? If not, you cannot legitimately calculate a plus four or large-sample confidence interval with only a 2% margin of error. Inference requires taking everything into consideration.

Notice also how large a sample size is required for a 2% or 3% margin of error. When dealing with proportions, the margin of error mm is always a value well below 1. Because mm appears in the denominator of the sample size equation and is squared, it drives the needed sample size upward. Studies of categorical variables (for example, clinical trials looking for improvement versus no improvement) tend to have rather large sample sizes compared with studies of quantitative variables.

APPLY YOUR KNOWLEDGE

19.10 Can you taste PTC? PTC is a substance that has a strong bitter taste for some people and is tasteless for others. The ability to taste PTC is inherited. Approximately 75% of all Italians can taste PTC, for example. You want to estimate the proportion of Americans who have at least one Italian grandparent and can taste PTC. Starting with the 75% estimate for Italians, how large a sample must you collect to estimate the proportion of PTC tasters within ±0.04± 0.04 with 90% confidence?

19.11 Planning a study of influenza. According to the CDC, each year in the United States, on average, 5% to 20% of the population gets the flu. Using 20% as our best guess (p*),(p^*), find how large a sample would be needed to get a ±2± 2 percentage point margin of error on a 95% confidence interval for the population proportion of Americans who get the flu in a given year.



James Cavallini/Science Source

HYPOTHESIS TESTS FOR A PROPORTION

We now turn to hypothesis tests. For the purpose of this discussion, we will assume that the conditions for inference we described earlier are met. When the null hypothesis H0: p=p0$H_0 : p = p_0$ is true, the sampling distribution of the sample proportion p^$\hat{p}$ is approximately Normal, centered on p0p_0 with standard deviation p0(1-p0)/n,$\sqrt{p_0 \left(1 - p_0\right) / n},$ as illustrated in Figure 19.3. Because H0H_0 gives us a hypothetical value for p,$p,$ the inaccuracy that plagues the large-sample confidence interval does not affect tests and, therefore, no adjustment is required.

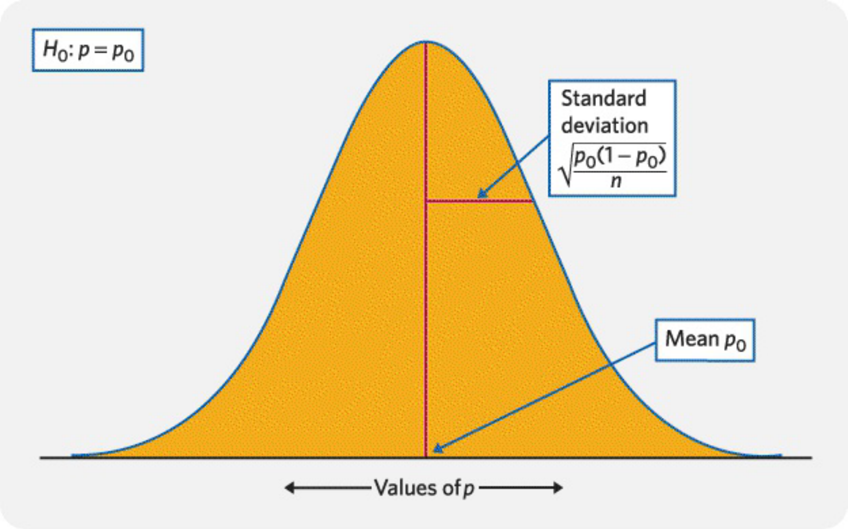
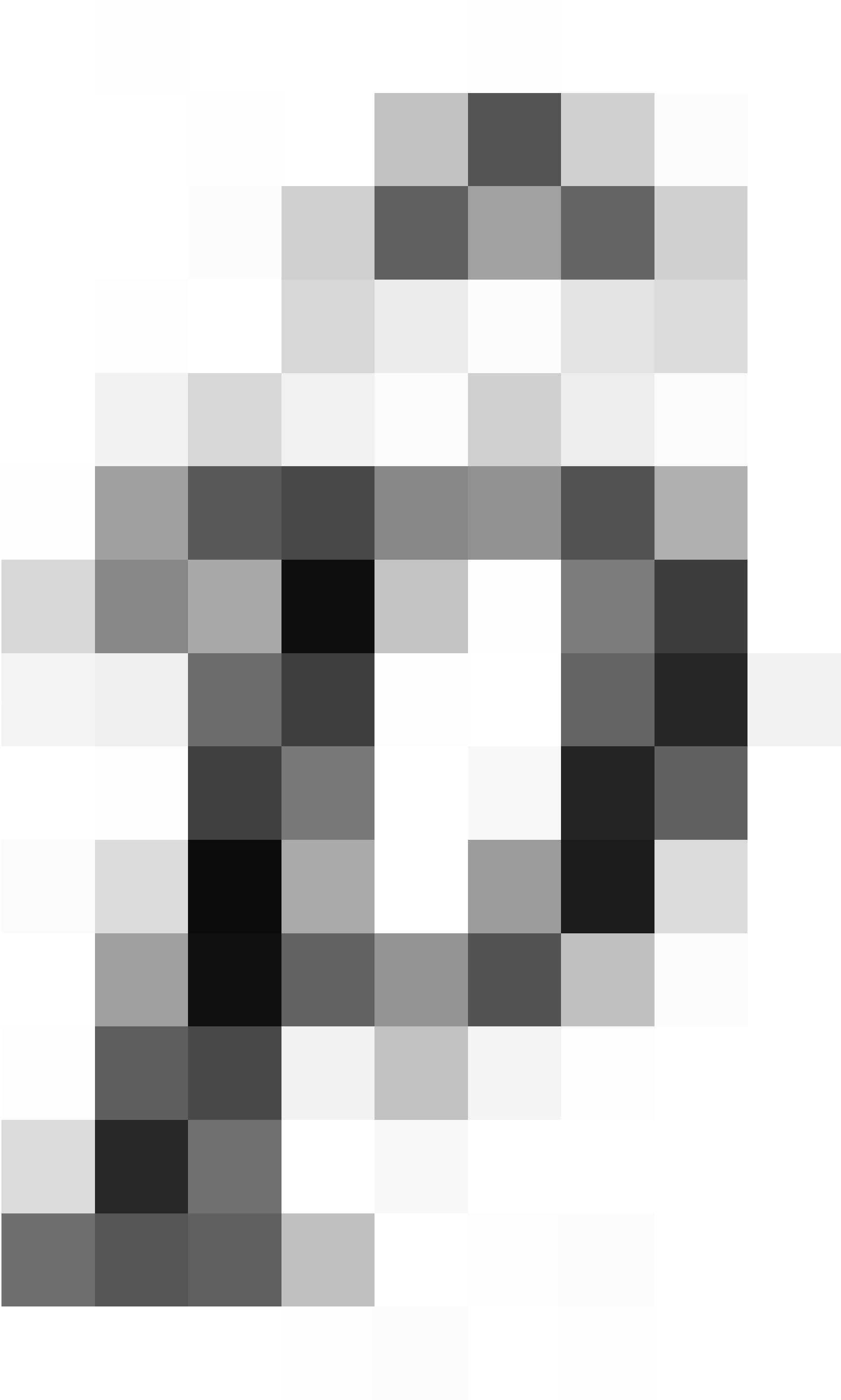


Figure 19.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 19.3 When H0: p=p0

$H_0 : p = p_0$ is true, the sampling

distribution of p^  is

approximately Normal with mean p0

p_0 and standard deviation

p0(1-p0)/n. $\sqrt{p_0\left(1-p_0\right)/n}.$

The test statistic for the null hypothesis H0: p=p0$H_0 : p = p_0$ is the sample proportion p^$\hat{p}$ standardized using the value p0p_0 specified by H0: $H_0 :$

z=p^-p0p0(1-p0)n

$$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0\left(1-p_0\right)}{n}}}$$

This zz statistic has approximately the standard Normal distribution when H0H_0 is true. As always, a test *P*-value is the probability of obtaining a test statistic at least as extreme (in the direction of the alternative) as that computed if the null hypothesis was true. Consequently, *P*-values for the z z test fora proportion come from the standard Normal distribution. Here is the procedure.

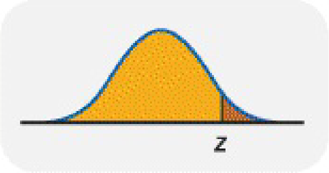
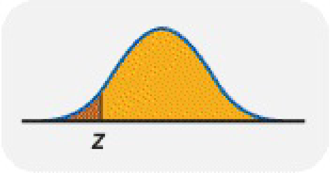
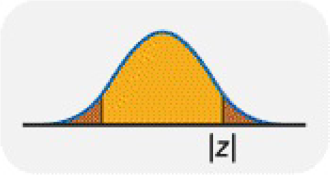
HYPOTHESIS TESTS FOR A PROPORTION

Draw an SRS of size nn from a large population that contains an unknown proportion pp of successes. To **test the hypothesis** H0: p=p0,$H_0 : p = p_0$, obtain the zz statistic:

z=p^-p0p0(1-p0)n

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

In terms of a variable ZZ having the standard Normal distribution, the approximate *P*-value for a test of H0H_0 against

Ha: p>p0$H_a : p > p_0$	isis	P(Z≥z)$P(Z \geq z)$	 Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company
Ha: p<p0$H_a : p < p_0$	isis	P(Z≤z)$P(Z \leq z)$	 Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company
Ha: p≠p0$H_a : p \neq p_0$	isis	2P(Z≥|z|)$2P(Z \geq \|z\|)$	 Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Technology with basic statistical inference capabilities will perform all necessary computations and provide both the test statistic and the *P*-value.

Use this test when the sample size nn is so large that both np0 np_0 and n(1-p0)$n(1-p_0)$ are 10 or greater.[12] They are the *expected* counts of successes and failures in nn observations when H0H_0 is true.

EXAMPLE 19.6 Do aphids right themselves when dropped?

The four-step process for any hypothesis test is outlined on page 365.

STATE: Pea aphids, *Acyrthosiphon pisum,* are small, wingless, sap-sucking insects that live on plants. They evade predators (such as ladybugs) by dropping off. A study examined the mechanism of aphid drops. Researchers hung live aphids upside down from delicate tweezers and then released them. The videotaped drops show that 19 of the 20 aphids landed on their legs.[13] Is this evidence that live aphids land right side up (on their legs) more often than chance alone would predict?

PLAN: Take pp to be the proportion of all live aphids that land on their legs when dropped. Aphids can land either on their legs or on their backs. If chance alone was at work, we would expect aphids to land on their legs half the time. Therefore, we test the hypotheses

H0: p=0.5Ha: p>0.5

$$\begin{aligned} H_0 &: p = 0.5 \\ H_a &: p > 0.5 \end{aligned}$$

SOLVE: The aphids in the study were reared in the lab and selected without bias. We can consider them to be the equivalent of an SRS. The expected counts of successes and failures under the null hypothesis are both 10 (np0=n(1-p0)=20×0.5=10).$(np_0 = n(1-p_0) = 20 \times 0.5 = 10)$. Therefore, the conditions for inference are met.

Next we use technology to obtain the sample statistic p^,$\hat{p}$, the test statistic z,z, and the one-sided *P*-value. We see from Figure 19.4 that the sample proportion of aphids landing on their legs is p^=0.95.$\hat{p} = 0.95$. The test statistic is z=4.02$z = 4.02$ and test *P*-value is 2.85E -5-5 (or simply reported as 0 because it is such a small value).

CONCLUDE: There is very strong evidence (P=0.00003)$(P = 0.00003)$ that more than half of live aphids dropped upside down from tweezers land on their legs, representing a better-than-chance success rate. This supports the idea that live aphids possess a mechanism for righting themselves when dropped (as cats do).

Nigel Cattlin/Science Source

Here are the computations performed by software for this example:

- Sample proportion: p^=19/20=0.95$\hat{p} = 19/20 = 0.95$
- Test statistic z:z :

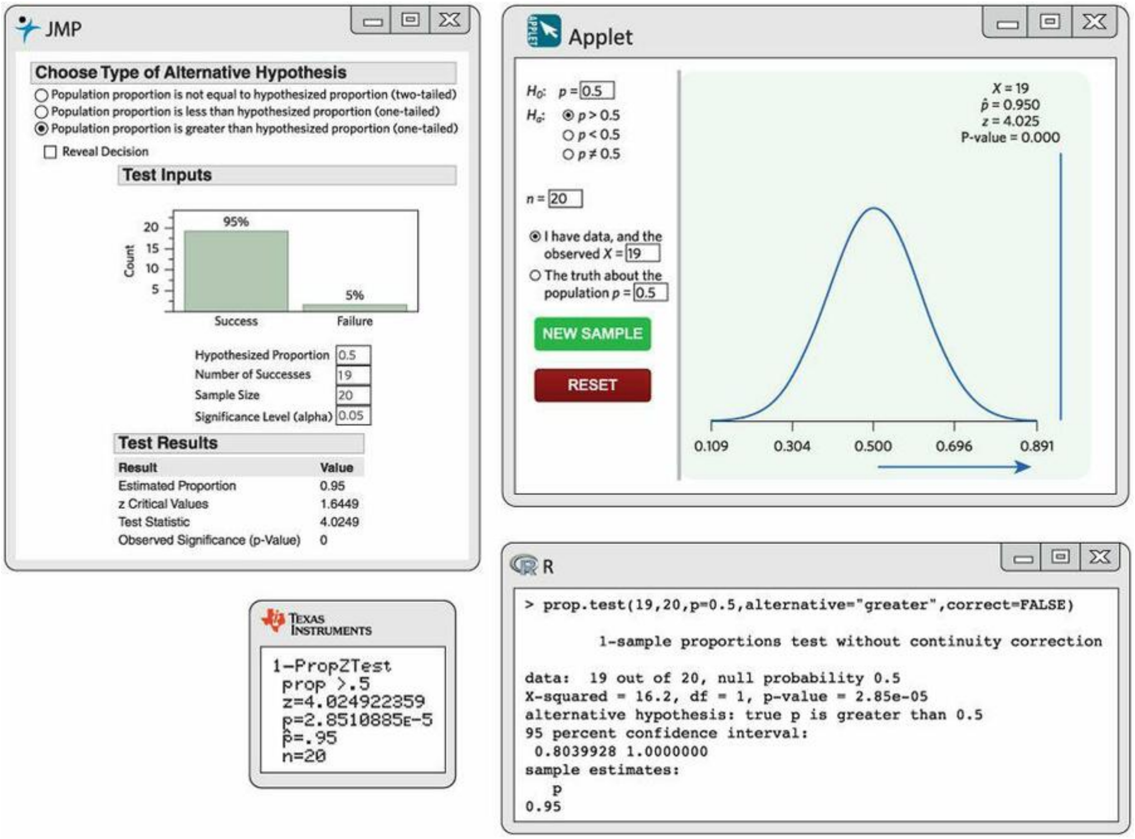
z=p^-p0p0(1-p0)n=0.95-0.5(0.5)(0.5)20=4.025

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.95 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{20}}} = 4.025$$

- Test *P*-value: This is the area to the right of z=4.025$z = 4.025$ under the standard Normal curve, because the alternative is one-sided in the greater-than direction. Software commands such as normalcdf(4.025,1E99,0,1) in the TI-83/84 calculator or =1-NORM.S.DIST(4.025,TRUE) in Excel give a *P*-value of 2.85E-5.-5. Table C tells us only that *P* is less than 0.0005 because the test statistic z=4.025$z = 4.025$ is greater than the largest critical value 3.291 listed for the zz distribution.

z*z^*	3.291
One-sided *P*	.0005

JMP, the TI-83 graphing calculator, and the applet in [Figure 19.4](#) all use the Normal approximation for the sampling distribution of p^.$\hat{p}$. The applet (*P-Value for a Test of One Proportion*, available on the book website) shows the value of p^=0.95 $\hat{p} = 0.95$ in relation to the sampling distribution of p^$\hat{p}$ when H0H_0 is true, helping us visualize why the test *P*-value is so small. Notice, however, that the R output does not provide the value of the test statistic z.z. Instead, it reports the value X-squared =16.2,$= 16.2$, which is the squared value of z=4.025.$z = 4.025$. When the categorical variable is binary ("success" or "failure"), this method is the equivalent of the zz test for a proportion. (This method is called the chi-square test for goodness of fit and we will examine it in more detail in [Chapter 21](#).) The *P*-value that R provides is the correct one-sided *P*-value matching the command alternative="greater". For any test, remember to always specify your null and alternative hypotheses before examining the sample data, and then obtain the one- or two-sided *P*-value that matches your alternative.



Choose Type of Alternative Hypothesis

○ Population proportion is not equal to hypothesized proportion (two-tailed)
○ Population proportion is less than hypothesized proportion (one-tailed)
● Population proportion is greater than hypothesized proportion (one-tailed)

☐ Reveal Decision

Test Inputs

Hypothesized Proportion 0.5
Number of Successes 19
Sample Size 20
Significance Level (alpha) 0.05

Test Results

Result	Value
Estimated Proportion	0.95
z Critical Values	1.6449
Test Statistic	4.0249
Observed Significance (p-Value)	0

H_0: p = 0.5
H_a: ● $p > 0.5$ ○ $p < 0.5$ ○ $p \neq 0.5$

n = 20

● I have data, and the observed X = 19
○ The truth about the population p = 0.5

$X = 19$
$\hat{p} = 0.950$
$z = 4.025$
P-value = 0.000

```
1-PropZTest
 prop >.5
 z=4.024922359
 p=2.8510885E-5
 p̂=.95
 n=20
```

```
> prop.test(19,20,p=0.5,alternative="greater",correct=FALSE)

        1-sample proportions test without continuity correction

data:  19 out of 20, null probability 0.5
X-squared = 16.2, df = 1, p-value = 2.85e-05
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.8039928 1.0000000
sample estimates:
   p
0.95
```

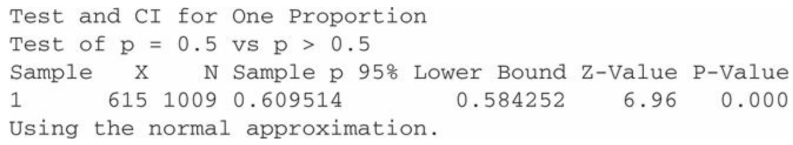
Figure 19.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 19.4 Output from a graphing calculator, two software packages, and an applet for

the hypothesis test about the righting behavior of aphids.

APPLY YOUR KNOWLEDGE

19.12 Soda consumption. "Majority of Americans say they try to avoid drinking soda": That's the title of a 2015 Gallup report based on a random sample of 1009 Americans interviewed about their eating habits. Here is the Minitab output for a z z test based on the survey findings:



```
Test and CI for One Proportion
Test of p = 0.5 vs p > 0.5
Sample    X     N Sample p 95% Lower Bound Z-Value P-Value
1       615  1009 0.609514        0.584252    6.96   0.000
Using the normal approximation.
```

a. What proportion of individuals in the sample said that they try to avoid drinking soda?

b. What are the null and alternative hypotheses corresponding to Gallup's title? Why is the alternative hypothesis one-sided?

c. What are the test statistic and *P*-value? Does the *P*-value support the claim made in the report's title? Explain your answer.

19.13 More on dropped aphids. The study in Example 19.6 also wondered whether the aphid aerial-righting behavior is an active or a passive mechanism (body shape and simple aerodynamics). So they used dead aphids in the same setting. Of the 23 dead aphids hung upside down from tweezers and then released, 12 landed on their legs.

a. Is this evidence that dead aphids land right side up (on their legs) more often than chance alone would predict? Follow the four-step process as illustrated in Example 19.6.

b. Write a short paragraph discussing aphid aerial righting behavior based on these results and those described in Example 19.6. Be sure to include sample statistics, test statistics, and *P*-values in your report.

19.14 Formulating the theory of inheritance. Gregor Mendel has been called the "father of genetics," having formulated laws of inheritance even before the function of DNA was discovered. In one famous experiment with peas, Mendel crossed pure breeds of plants that produced either smooth peas or wrinkled peas. The first-generation peas were all smooth. In the second generation (F2), he obtained 5474 smooth peas and 1850 wrinkled peas. Do these data agree with the conclusion of 75% dominant-trait (in this case "smooth") occurrence in F2? Follow the four-step process as illustrated in Example 19.6.

19.15 No zz test. Explain why we can't use the zz test for a proportion in the following situations:

a. You want to know if people have predictive psychic abilities when answering their phones. You ask 10 subjects to guess which of 5 possible friends is calling them before they pick up the phone to test the hypotheses H0: p=0.2$H_0 : p = 0.2$ (no psychic ability) versus Ha: p>0.2.$H_a : p > 0.2.$

b. You bet a friend that 90% or more of the students taking the Bio 101 class know their blood types. You contact an SRS of 80 of the 1211 students taking Bio 101 this term to test the hypotheses H0: p=0.9$H_0 : p = 0.9$ versus Ha: p>0.9.$H_a : p > 0.9.$

CHAPTER 19 SUMMARY

- When the data are an SRS of size n,n, tests and confidence intervals for a population proportion pp are based on the **sample proportion** p^.$\hat{p}$.
- When nn is large, p^$\hat{p}$ has approximately the Normal distribution with mean pp and standard deviation p(1-p)/n.$\sqrt{p\,(1-p)\,/n}$.
- Technology can be used to automate the Normal computations required to obtain a confidence interval for the proportion pp of a population or to test a set of null and alternative hypotheses about p.p.
- A level CC **large-sample confidence interval for** pp is

  p^±z*p^(1-p^)n

  $$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

  where z*z^* is the critical value for the standard Normal curve with area CC between -z*$-z^*$ and z*.z^*.
- The true confidence level of the large-sample interval can be substantially less than the planned level CC unless the sample is very large. *We recommend using the plus four interval instead.*
- To obtain a more accurate confidence interval, add four imaginary observations, two successes and two failures, to your sample. Then use the same formula for the confidence interval. This is the **plus four confidence interval**. Use this interval in practice for a confidence level of 90% or higher and a sample size nn of at least 10.
- The **sample size** needed to obtain a confidence interval with an approximate margin of error mm for a population proportion is

  n=z*m2p*(1-p*)

  $$n = \left(\frac{z^*}{m}\right)^2 p^*\,(1-p^*)$$

  where p*p^* is a guessed value for the sample proportion p^$\hat{p}$ and z*z^* is the standard Normal critical point for the level of confidence you want. If you use $p^* = 0.5$ in this formula, the margin of error of the interval will be less than or equal to m no matter what the value of $\hat{p}$ is.

- A **hypothesis test for** $H_0: p = p_0$ is based on the value of the z statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

when H_0 is true. The test *P*-value, computed as an area under the standard Normal distribution, tells us whether to reject or fail to reject H_0. Use this test in practice when $np_0 \geq 10$ and $n(1-p_0) \geq 10$.

THIS CHAPTER IN CONTEXT

Over the course of Chapters 19 through 22, we describe statistical inference procedures used when the response variable is categorical. That is, for each individual or observation in the sample(s), we record an attribute (such as gender or opinion) rather than a numerical value, as explained in Chapter 1. Ultimately, we are interested in the proportion pp of successes in one or more populations.

In Chapter 19 we compute confidence intervals and test hypotheses about one population proportion p,$p,$ just as we computed confidence intervals and tested hypotheses about a population mean μμ in Chapter 17 when the response variable was quantitative. In Chapter 20 we describe inference procedures that let us compare the proportions of successes in two populations when the data have been collected using two independent samples.

The computations described in this chapter build on the Normal calculations learned in Chapter 11 and on the properties of the sampling distribution of p^$\hat{p}$ introduced in Chapter 13. These Normal calculations are approximately correct only when the counts of successes and failures in the sample (observed or expected from a null hypothesis) are large enough. For this reason, you must check that these count assumptions are met before performing the inference procedure you want.

CHECK YOUR SKILLS

A CDC report on secondhand smoke at home gives the following 95% confidence interval for the percent of California households that are free of secondhand smoke: (90.8, 92.2). Use this information to answer Exercises 19.16 to 19.18.

19.16 The correct interpretation for this confidence interval is that we can be 95% confident that

a. the proportion of households free of secondhand smoke in another sample of California households would be between 0.908 and 0.922.
b. the population mean number of households in California that are free of secondhand smoke is between 90.8 and 92.2.
c. the true proportion of all California households that are free of secondhand smoke is between 0.908 and 0.922.

19.17 What is the margin of error for this interval?

a. 0.007
b. 0.014
c. 0.028

19.18 A 90% confidence interval based on the same data would have

a. a smaller margin of error.
b. the same margin of error.
c. a larger margin of error.

A study examined the effectiveness of an ear-molding technique used to correct ear deformities in newborns without the need for surgery. A plastic mold (the EarWell Infant Ear Correction System) was placed around the deformed ear shortly after birth for about two weeks. The study reports that 152 of 158 procedures were successful.[14] *Use this information to* *answer Exercises 19.19 to 19.21.*

19.19 What is the sample proportion p^$\hat{p}$ of successful procedures?

a. 1.52
b. 0.962
c. 0.06

19.20 We would like to estimate with 95% confidence the true proportion of successful ear-molding procedures among all newborns needing ear-deformity correction. We can use

a. either the large-sample method or the plus four method because the sample size is large.
b. the plus four method but not the large-sample method because the counts of successes and failures in the sample are too small but the sample size is large enough.
c. neither method because the sample size is too small.

19.21 What is the plus four 95% confidence interval for the true proportion of successful ear-molding procedures among all newborns needing ear-deformity correction?

a. (152, 158)
b. (0.932, 0.992)
c. (0.917, 0.984)

19.22 (Optional) How many observations must be recorded to estimate a population with unknown proportion pp to within ±0.02± 0.02 with 95% confidence? Use 0.5 as the conservative guess for p.p.

a. n=25$n = 25$
b. n=1225$n = 1225$
c. n=2401$n = 2401$

A farmer wants to test the effectiveness of a pest control method in allowing strawberry blooms to yield marketable strawberries. Out of a random sample of 100 blooms, 77 yield marketable strawberries. Is this evidence that more than two-thirds of all blooms grown with this pest *control method end up as marketable strawberries? Use this information to answer Exercises 19.23 to 19.25.*

19.23 What is the null hypothesis for a test to answer this question?

a. $H0: p=0.5$$H_0 : p = 0.5$
b. $H0: p=0.667$$H_0 : p = 0.667$
c. $H0: p=0.77$$H_0 : p = 0.77$

19.24 The alternative hypothesis for a test to answer this question is

a. one-sided in the greater-than direction.
b. one-sided in the less-than direction.
c. two-sided.

19.25 The value of the statistic for this test is $z=2.192$.$z = 2.192$. What is the approximate *P*-value for this test?

a. 0.0140.014
b. 0.0290.029
c. 0.0470.047

CHAPTER 19 EXERCISES

We recommend using the plus four method for all confidence intervals for a proportion. However, the large-sample method is acceptable when the guidelines for its use are met.

19.26 Health care–associated infections. Eliminating infections acquired during hospitalization is a health care priority. Researchers aiming to quantify the scale of this problem in the United States reported their survey findings as follows:[15] "Of 11,282 patients, 452 had 1 or more health care–associated infections (4.0%; 95% confidence interval, 3.7 to 4.4)." Interpret this confidence interval in context.

19.27 White-eyed fruit flies. Wild-type fruit flies have red eyes, but a recessive mutation produces white-eyed individuals. A researcher wants to assess the frequency of heterozygous individuals among red-eyed fruit flies. A heterozygous red-eyed fly crossed with a white-eyed mutant will have mixed progeny. Of the 100 red-eyed fruit flies crossed with white-eyed mutants, 11 produced mixed progeny.

a. Show that the conditions for the large-sample confidence interval are not met. Show that the conditions for the plus four interval are met.

b. Give the plus four 95% confidence interval for the proportion of heterozygotes in the population of red-eyed fruit flies studied.

19.28 Handling contact lenses. Failure to follow recommended contact lens wear and care practices can lead to serious eye infection. A survey of a random sample of 281 Americans who wear contact lenses regularly asked about contact lens practices. The survey found that 139 (49.5%) respondents do not consistently wash their hands before handing their contact lenses.[16]

a. Obtain a plus four 99% confidence interval for the proportion p p of all American contact lens wearers who do not consistently wash their hands before handing their lenses.

b. Verify that the conditions for your confidence interval are met.

c. The researchers indicated that this survey had a substantial nonresponse rate. How does this information affect your interpretation of the confidence interval in context?

d. Survey participants simply answered a questionnaire, and no attempt was made to verify the answers. How does this information affect your interpretation of the confidence interval in context?

19.29 White-eyed fruit flies, continued (optional). Go back to the experiment of Exercise 19.27. How many red-eyed fruit flies should be used in this experiment to obtain a 95% confidence interval with a margin of error no more than 3%? Use 20% as a conservative (worst-case) estimate of the percent of heterozygotes in the population.

19.30 Color blindness (optional). Color blindness, or dyschromatopsia, is a form of genetic deficiency in color perception. The condition is much more prevalent among males than females, pointing to a genetic connection with the X chromosome. The frequency of dyschromatopsia in the Caucasian American male population is roughly 8%, but this proportion might be different among males of other ethnicities. We want to estimate the proportion of Asian American males who are color-blind. How large a sample size do we need to obtain a 95% confidence interval with a margin of error no greater than 2%, or 0.02? Use 0.1 for p*.p^*.

19.31 The lingering smell of asparagus. The consumption of asparagus results in the release of methanethiol and *S*-methyl thioesters metabolites in the urine. To some people, these metabolites have a very distinctive odor, but others cannot detect the odor (“asparagus anosmia”). How common is asparagus anosmia? In a genetic study aiming to find the genetic markers supporting the phenotype, researchers contacted 6909 participants in a large scientific cohort and found that 4161 had asparagus anosmia.[17] Assuming that the cohort is representative of the American adult population, obtain a

90% confidence interval for the true proportion of Americans with asparagus anosmia.

19.32 Side effects. An experiment on the side effects of pain relievers assigned patients with arthritis to take one of several over-the-counter pain medications. Of the 440 patients who took one brand of pain reliever, 23 suffered some "adverse symptom."

a. If 10% of all patients suffer adverse symptoms, what would be the sampling distribution of the proportion with adverse symptoms in a sample of 440 patients?

b. Does the experiment provide strong evidence that fewer than 10% of patients who take this medication have adverse symptoms? Verify that the conditions for inference are met. State the hypotheses, obtain the test statistic and *P*-value, and conclude in context.

19.33 Cancer-detecting dogs. A study was designed to determine whether dogs can be trained to identify urine specimens from individuals with bladder cancer. Dogs were first trained to discriminate between urine specimens from patients with bladder cancer and urine specimens from patients with other conditions. After the training was completed, the dogs had to pick one of seven new urine specimens. Each time, only one of the seven urine specimens came from a patient with bladder cancer. Out of 54 trials, the dogs identified the correct urine specimen 22 times.[18]

a. If the dogs were simply picking a urine specimen at random, we would expect them to be correct, on average, 1 out of 7 times. The experiment was designed to test whether dogs can perform better than chance. State the null and alternative hypotheses for this test.

b. Obtain the test statistic and the *P*-value. What do you conclude?

19.34 Matched pairs. One-sample procedures for proportions, like those for means, are used to analyze data from matched pairs designs. Here is an example. Each of 50 subjects tastes two unmarked cups of coffee and says which he or she prefers. One cup in each pair contains instant coffee; the other, fresh-brewed coffee. Thirty-one of the subjects prefer the fresh-brewed coffee. Take p to be the proportion of the population who would prefer fresh-brewed coffee in a blind tasting.

a. Test the claim that a majority of people prefer the taste of fresh-brewed coffee. State the hypotheses and report the z statistic and its *P*-value. Is your result significant at the 5% level? What is your practical conclusion?
b. Find a 90% confidence interval for p.
c. When you do an experiment like this, in what order should you present the two cups of coffee to the subjects?

In responding to Exercises 19.35 through 19.41, follow the **Plan, Solve,** *and* **Conclude** *steps of the four-step process. It may be helpful to restate in your own words the* **State** *information given in the exercise.*

19.35 Lead in tap water. The U.S. Environmental Protection Agency (EPA) defines the Action Level for lead in tap water as 15 parts per billion (ppb; $15\ \text{ppb} = 0.015\ \text{mg}/\text{l}$). This is a level above which further analysis and public education campaigns must be done, but not yet a level causing a real health risk. A random sample of 10,341 water specimens from the Washington, D.C. area found that 7496 water specimens had lead concentrations up to 15 ppb.[19] What is the proportion of all possible water specimens in D.C. that would exceed the lead Action Level? Use a 95% confidence level.

19.36 Vote for the best face? We often judge other people by their faces. It appears that some people even judge candidates for elected office by their faces. Researchers showed head-and-shoulders photos of the two main candidates in 32 races for the U.S. Senate to many subjects (dropping subjects who recognized one of the candidates) to see which candidate was rated "more competent" based solely on the photos. On Election Day, the candidates whose faces looked more competent won 22 of the 32 contests.[20] If faces don't influence voting, half of all races in the long run should be won by the candidate with the better face. Is there evidence that the candidate with the better face wins more than half the time? Follow the four-step process as illustrated in Example 19.6.

19.37 Depression. The Gallup-Healthways Well-Being Index is a comprehensive survey of the health status of Americans. In 2009, the survey asked a random sample of 258,141 adults, “Have you ever been told by a physician or a nurse that you have depression?” Of these respondents, 43,884 answered “yes.” Estimate with 95% confidence the proportion of adults in America who have ever been told by a health professional that they have depression.

19.38 The polarizing taste of cilantro. The culinary herb cilantro, *Coriandrum sativum,* is very polarizing: Some people love it and others hate it. A genetic component is suspected to be at play. A survey of 12,087 American adults of European ancestry asked whether they like or dislike the taste of cilantro. A total of 3181 said that they dislike the taste of cilantro.[21] Estimate with 95% confidence the proportion of all American adults of European ancestry who dislike the taste of cilantro.

19.39 Children’s food choices. Do popular cartoon characters on food packages influence children’s food choices? A study asked 40 young children (ages four to six) to taste Graham Crackers presented in a package either with or without a popular cartoon character. When asked to indicate which of the two options they would prefer to eat for a snack, 35 chose the version with a cartoon on the package.[22] Is this evidence that young children prefer snacks packaged with popular cartoon images more than chance alone would explain? Estimate with 95% confidence the proportion of young children who prefer Graham Crackers packaged with popular cartoon images. Answer both questions with appropriate inference methods.

19.40 The polarizing taste of cilantro, continued. One reason sometimes given for the dislike of cilantro is that it tastes like soap. The survey in Exercise 19.38 asked 14,604 American adults of European ancestry whether cilantro tastes soapy to them; a total of 1994 respondents answered that it does. Estimate with 95% confidence the proportion of all American adults of European ancestry who find that cilantro tastes soapy.

19.41 Staph infections. A study investigated ways to prevent staph infections in surgery patients. In a first step, the researchers examined the nasal secretions of a random sample of 6771 patients admitted to various hospitals for surgery. They found that 1251 of these patients tested positive for *Staphylococcus aureus*, the bacterium responsible for most staph infections.[23] Calculate a 95% confidence interval for the proportion of all patients admitted for surgery who are carriers of *S. aureus*.

Exercises marked with the Large Data Set *icon guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. You can find many more such exercises in the book's review chapters (Chapters 8, 16, and 25). Note that short answers are not available for these exercises to enable their use for class work or assignments.*

19.42 Estimating body fat in men. In Exercises 8.16 and 8.17 (page 208), and 17.46 (page 443) you examined the data from the data file *Large.Bodyfat*, which provides the percent of body fat, age, weight, height, and 10 body circumference measurements for 252 adult men.

a. Use the columns for weight (in pounds) and height (in inches) to compute a new variable representing an individual's body mass index (BMI). The formula for BMI (when using pounds and inches) is

BMI=703×weightheight2

$$BMI = \frac{703 \times weight}{height^2}$$

b. Obesity is defined as a BMI of 30.0 or greater. Find the proportion of individuals in this sample who are considered obese based on their BMI. Make sure to exclude the individual with an incorrect height value that you identified in Exercise 8.16.

c. The data are a random sample of 252 adult males taken in 1985. Give a 95% confidence interval for the proportion of obese adult males in 1985.

d. The most recent data from the CDC indicate that, currently, 28.4% of adult men are obese.[24] How do the 1985 data compare with the current estimate? Use your results from part c to answer this question.

CHAPTER 20 Comparing Two Proportions

Doug Pensinger/Getty Images

IN THIS CHAPTER WE COVER...

- Two-sample problems: proportions
- The sampling distribution of a difference between proportions
- Large-sample confidence intervals for comparing proportions
- Accurate confidence intervals for comparing proportions
- Hypothesis tests for comparing proportions
- Relative risk and odds ratio
- *Discussion: Assessing and understanding health risks*

In a **two-sample problem**, we want to compare two populations or the responses to two treatments based on two independent samples. When the comparison involves the *means* of two populations, we use the two-sample *t* methods of Chapter 18. Now we turn to methods to compare the *proportions* of successes in two populations when the data are categorical.

TWO-SAMPLE PROBLEMS: PROPORTIONS

We will use notation similar to that used in our study of two-sample *t* statistics. The groups we want to compare are Population 1 and Population 2. We have a separate SRS from each population or responses from two treatments in a randomized comparative experiment. A subscript shows which group a parameter or statistic describes. Here is our notation:

Population	**Population proportion**	**Sample size**	**Sample proportion**
1	p1p_1	n1n_1	p^1$\hat{p}_1$
2	p2p_2	n2n_2	p^2$\hat{p}_2$

We compare the populations by doing inference about the difference p1-p2 $p_1 - p_2$ between the population proportions. The statistic that estimates this difference is the difference between the two sample proportions, p^1-p^2. $\hat{p}_1 - \hat{p}_2$.

EXAMPLE 20.1 How to quit smoking

STATE: If you want to quit smoking, is it better to stop abruptly or gradually? Researchers in England recruited adults who smoked at least 15 cigarettes daily and were willing to quit smoking. The subjects were randomly assigned to either the abrupt-cessation group or the gradual-cessation group (who reduced the amount smoked two weeks before stopping). Both groups received similar support in the form of behavioral counseling and nicotine patches. Verified abstinence was achieved four weeks later by 174 of the 355 subjects in the abrupt-cessation group and 134 of the 342 subjects in the gradual-cessation group.[1] How much more likely are smokers to still abstain from smoking four weeks after they stopped if they quit abruptly rather than gradually?

PLAN: Take smokers stopping abruptly to be Population 1 and smokers stopping gradually to be Population 2. The population proportions who still abstain from smoking four weeks after stopping are p1p_1 for the abrupt condition and p2p_2 for the gradual condition. We want to estimate the difference p1-p2$p_1 - p_2$ with 95% confidence.

vchal/Shutterstock.com

SOLVE: Inference about population proportions is based on the sample proportions

p^1=174355=0.4901 (abrupt)p^2=134342=0.3918 (gradual)

$$\begin{aligned}\hat{p}_1 &= \frac{174}{355} = 0.4901 \quad \text{(abrupt)} \\ \hat{p}_2 &= \frac{134}{342} = 0.3918 \quad \text{(gradual)}\end{aligned}$$

We see that 49% of the smokers in the study who stopped abruptly but only 39% of those who stopped gradually still abstained from smoking four weeks after stopping. We want to estimate p1-p2$p_1 - p_2$ with the difference between the sample proportions from this experiment:

p^1-p^2=0.4901-0.3918=0.0983

$$\hat{p}_1 - \hat{p}_2 = 0.4901 - 0.3918 = 0.0983$$

To complete the Solve step, we must know how differences between sample proportions behave.

THE SAMPLING DISTRIBUTION OF A DIFFERENCE BETWEEN PROPORTIONS

To use p^1-p^2$\hat{p}_1 - \hat{p}_2$ for inference, we must know its sampling distribution. In Chapter 18 we saw that when two random variables are Normally distributed, the new variable “difference” also follows a Normal distribution, centered on the difference between the two variables’ means and with variance equal to the sum of the two variables’ variances. In Chapter 19 we learned that the sampling distribution of p^$\hat{p}$ is approximately Normal when the sample size nn is large, with mean and standard deviation pp and p(1-p)/n,$\sqrt{p(1-p)/n}$, respectively. Here are the facts on which we rely when conducting inference for the difference between two population proportions:

- When the samples are large, the distribution of p^1-p^2$\hat{p}_1 - \hat{p}_2$ is **approximately Normal**.
- The **mean** of the sampling distribution of p^1-p^2$\hat{p}_1 - \hat{p}_2$ is p1-p2. $p_1 - p_2$. That is, the difference between sample proportions is an unbiased estimator of the difference between population proportions.
- The **standard deviation** of the distribution is

p1(1-p1)n1+p2(1-p2)n2

$$\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$$

Figure 20.1 displays the sampling distribution of p^1-p^2.$\hat{p}_1 - \hat{p}_2$. The standard deviation of p^1-p^2$\hat{p}_1 - \hat{p}_2$ involves the unknown parameters p1 p_1 and p2.p_2. Just as in the previous chapter, we must replace these by estimates in order to do inference. Also, just as in the previous chapter, we do this a bit differently for confidence intervals and for tests.

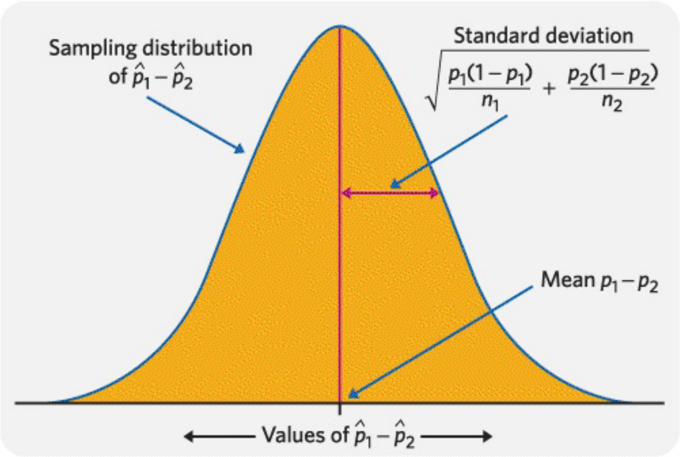
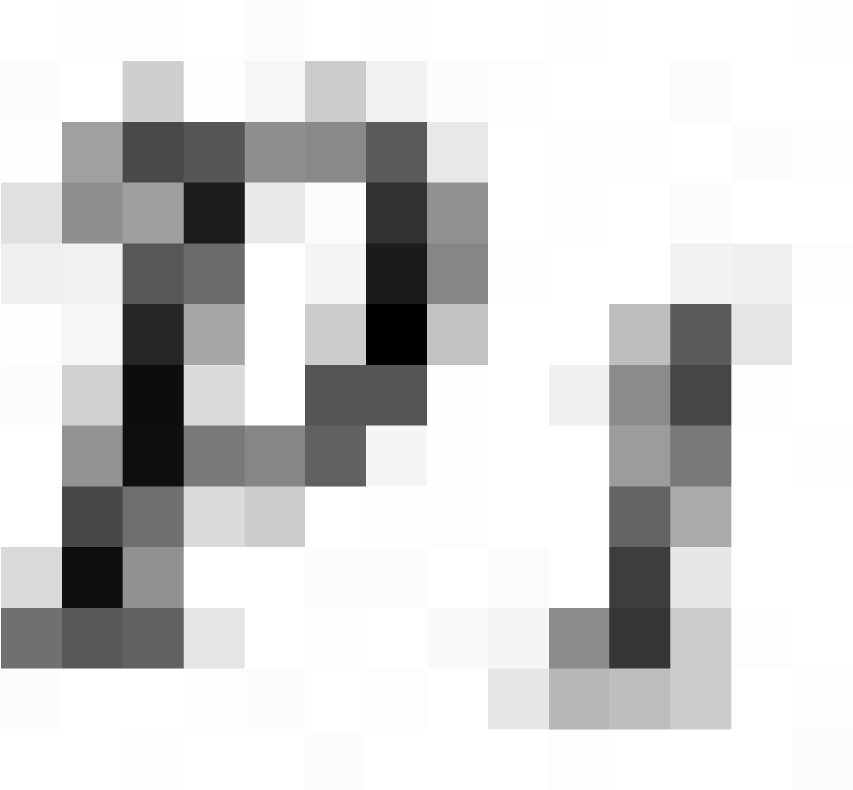


Figure 20.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 20.1 Select independent SRSs from two populations having proportions of successes p1 and p2.



The proportions of successes in the two samples are p^1

and p^2

$\hat{p}_2$ When the samples are large, the sampling distribution of the difference p^1-p^2 $\hat{p}_1 - \hat{p}_2$ is appropriately Normal.

LARGE-SAMPLE CONFIDENCE INTERVALS FOR COMPARING PROPORTIONS

To obtain a confidence interval we can actually compute, one option is to replace the population proportions p1p_1 and p2p_2 in the standard deviation of the sampling distribution with the two sample proportions. The result is the **standard error** of the statistic p^1-p^2:$\hat{p}_1 - \hat{p}_2$:

SE=p^1(1-p^1)n1+p^2(1-p^2)n2

$$\text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

The confidence interval has the same form we saw in the previous chapter:

estimate±z*SEestimate

$$\text{estimate} \pm z^* \text{SE}_{\text{estimate}}$$

LARGE-SAMPLE CONFIDENCE INTERVAL FOR COMPARING TWO PROPORTIONS

Draw an SRS of size n1n_1 from a large population having proportion p1p_1 of successes, and draw an independent SRS of size n2n_2 from another large population having proportion p2p_2 of successes. When n1n_1 and n2n_2 are large, an approximate level CC **confidence interval for** p1-p2$p_1 - p_2$ is

(p^1-p^2)±z*SE

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \text{SE}$$

In this formula the standard error SE of p^1-p^2$\hat{p}_1 - \hat{p}_2$ is

SE=p^1(1-p^1)n1+p^2(1-p^2)n2

$$\text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and z*z^* is the critical value for the standard Normal density curve with area CC between -z*$-z^*$ and z*.z^*. Technology with basic statistical inference capabilities will perform all necessary computations and produce the lower and upper boundaries of the confidence interval.

Use this interval only when the number of successes and the number of failures are each 10 or more in both samples.

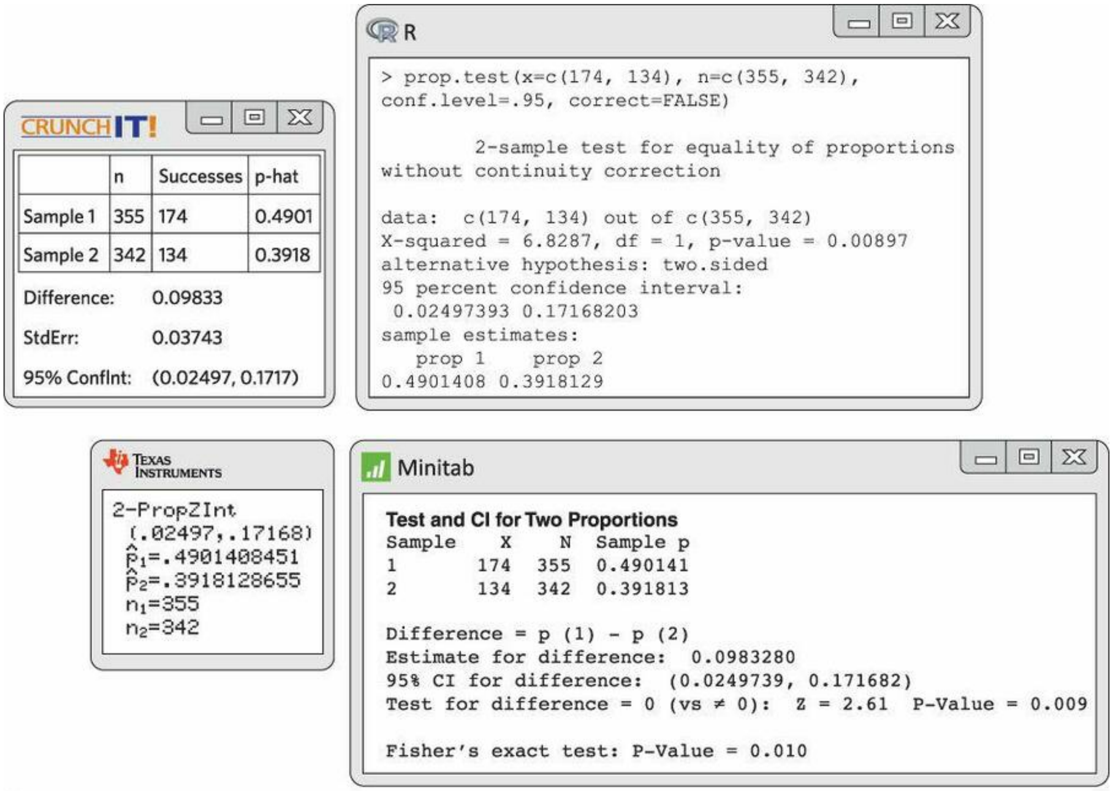
EXAMPLE 20.2 How to quit smoking, continued

We can now complete Example 20.1. Here is a summary of the basic information:

Population	Population description	Sample size	Number of successes	Sample proportion
1	abrupt	n1=355$n_1 = 355$	174	p^1=174/355=0.4901$\hat{p}_1 = 174/355 = 0.4901$
2	gradual	n2=342$n_2 = 342$	134	p^2=134/342=0.3918$\hat{p}_2 = 134/342 = 0.3918$

SOLVE: We will give a 95% confidence interval for p1-p2,$p_1 - p_2$, the difference between the proportions still abstaining from smoking after four weeks for the two smoking cessation approaches. The experiment randomly assigned the subjects to the conditions, so we have two independent random samples. To check that the large-sample confidence interval is safe, look at the counts of successes and failures in the two samples. All of these four counts are much larger than 10, so the large-sample method will be accurate.

Using technology, as shown in Figure 20.2, we obtain the confidence interval 0.0250 to 0.1717 for p1-p2.$p_1 - p_2$.



	n	Successes	p-hat
Sample 1	355	174	0.4901
Sample 2	342	134	0.3918

Difference: 0.09833

StdErr: 0.03743

95% ConfInt: (0.02497, 0.1717)

```
> prop.test(x=c(174, 134), n=c(355, 342),
conf.level=.95, correct=FALSE)

        2-sample test for equality of proportions
without continuity correction

data:  c(174, 134) out of c(355, 342)
X-squared = 6.8287, df = 1, p-value = 0.00897
alternative hypothesis: two.sided
95 percent confidence interval:
 0.02497393 0.17168203
sample estimates:
   prop 1    prop 2
0.4901408 0.3918129
```

```
2-PropZInt
 (.02497,.17168)
 p̂1=.4901408451
 p̂2=.3918128655
 n1=355
 n2=342
```

```
Test and CI for Two Proportions
Sample     X    N  Sample p
1        174  355  0.490141
2        134  342  0.391813

Difference = p (1) - p (2)
Estimate for difference:  0.0983280
95% CI for difference:  (0.0249739, 0.171682)
Test for difference = 0 (vs ≠ 0):  Z = 2.61  P-Value = 0.009

Fisher's exact test: P-Value = 0.010
```

Figure 20.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 20.2 Output from a graphing calculator and three statistical programs for a study comparing the success rates of abrupt and gradual smoking cessation.

CONCLUDE: We are 95% confident that the percent of smokers provided with behavioral support and nicotine patch who still abstain four weeks after quitting is somewhere between 2.5 and 17.2 percentage points greater when quitting abruptly than when quitting gradually.

The computations performed by software in this example involve the following steps (with some roundoff error):

- Sample estimate: p^1-p^2=0.4901-0.3918=0.0983

  $\hat{p}_1 - \hat{p}_2 = 0.4901 - 0.3918 = 0.0983$

- Standard error of the estimate:

SE=p^1(1-p^1)n1+p^2(1-p^2)n2=(0.4901)(0.5099)355+(0.3918)(0.6082)342=0.001401=0.03743

$$\begin{aligned} \text{SE} &= \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ &= \sqrt{\frac{(0.4901)(0.5099)}{355}+\frac{(0.3918)(0.6082)}{342}} \\ &= \sqrt{0.001401} = 0.03743 \end{aligned}$$

- Standard Normal critical value for a 95% confidence level: z*=1.960 $z^* = 1.960$ (rounded to three decimal places, obtained from software or Table C)
- Margin of error: m=z*SE=1.960(0.03743)=0.0734 $m = z^*\text{SE} = 1.960\,(0.03743) = 0.0734$
- 95% confidence interval for p1-p2: $p_1 - p_2$:

(p^1-p^2)±m=0.0983±0.0734, or 0.0249 to 0.1717

$$(\hat{p}_1 - \hat{p}_2) \pm m = 0.0983 \pm 0.0734, \text{ or } 0.0249 \text{ to } 0.1717$$

Figure 20.2 displays the output for Example 20.2 from a graphing calculator and three statistical software programs. All provide the sample statistics p^1 $\hat{p}_1$ and p^2 $\hat{p}_2$ as well as the confidence interval for p1-p2. $p_1 - p_2$. Minitab and R provide additional information related to a two-sided hypothesis test to complement the confidence interval. We will discuss these outputs later in the chapter. Not all software packages provide a simple way to obtain confidence intervals for proportions; Excel, for instance, has no such feature.

Like the large-sample confidence interval for a single proportion p p, *the large-sample interval for* p1-p2 $p_1 - p_2$ *generally has a true confidence level less than the level you asked for.* This inaccuracy is not as serious as in the one-sample case, however, at least if our guidelines for use are followed. In the next section we see how the plus four method described in Chapter 19 can be used for two-sample situations.

CAUTION

Because the z z methods of Chapters 19 and 20 offer only approximate level C C confidence intervals for proportions, a number of approaches to improve on them exist (such as the plus four method), which also means that knowing which approach each statistical program uses can be a bit challenging. For example, JMP and R (unless you specify correct=FALSE, as we did) give slightly different intervals. *It is always a good idea to consult*

CAUTION

your technology's help documentation if you need clarification.

The confidence interval we obtained should be interpreted in the context of the problem, thereby going beyond its simple statistical definition. Overall, the success rates after four weeks were not impressive for either group (only 49% and 39% in the study for the abrupt and gradual conditions, respectively). The confidence interval we obtained tells us that the true difference between the success rates in the two populations could be as small and nearly trivial as 2.5 percentage points, or it could be as much as a substantial 17.2 percentage points. As we discussed in Chapter 19 in the section on sample size (page 484), inference for proportions typically requires really large sample sizes to have reasonably small margins of error. Even with nearly 700 participants, the wide confidence interval produced by the study makes it difficult to know whether quitting abruptly has a substantially higher chance of being successful at four weeks after quitting. We should also consider that smokers typically want to quit for life, so a four-week assessment may not be particularly useful. Exercise 20.1 compares the percent of participants who still abstained six months after quitting, a better—though still imperfect—evaluation of long-term success.

APPLY YOUR KNOWLEDGE

20.1 More about smoking cessation. The secondary outcome of the experiment described in Example 20.1 was smoking abstinence six months after quitting. Verified abstinence was achieved at that point by 78 of the 355 subjects in the abrupt-cessation group and 53 of the 342 subjects in the gradual-cessation group. Give a 95% confidence interval for the difference between the proportions of smokers who still abstain from smoking six months after they stopped, when quitting either abruptly or gradually. Follow the four-step process as illustrated in Examples 20.1 and 20.2.

20.2 Botox for excessive sweating. Hyperhidrosis is a stressful medical condition characterized by chronic excessive sweating. Primary hyperhidrosis is inherited and typically starts during adolescence. Botox (botulinum toxin) was approved for treatment of hyperhidrosis in adults in 2004. A clinical trial examined the effectiveness of Botox for excessive armpit sweating in teenagers aged 12 to 17 years. Participants received either a Botox injection or a placebo injection (a simple saline solution) and were examined four weeks later to see if their sweating had been reduced by 50% or more. Here are the study's findings:[2]

	At least 50% reduction		
Treatment	**Yes**	**No**	**Total**
Botox	84	20	104
Placebo	44	64	108
Total	128	84	212

PYMCA/SuperStock

Give a 90% confidence interval for the difference (Botox minus placebo) in the proportions of teenagers with hyperhidrosis who experience at least 50% reduction in sweating. Follow the four-step process as illustrated in Examples 20.1 and 20.2.

ACCURATE CONFIDENCE INTERVALS FOR COMPARING PROPORTIONS

As in Chapter 19, another option for obtaining a confidence interval for p1-p2$p_1 - p_2$ is to add four imaginary observations to the sample data. This simple adjustment has been shown to greatly improve the accuracy of the confidence interval compared with the large-sample method.[3]

PLUS FOUR CONFIDENCE INTERVAL FOR COMPARING TWO PROPORTIONS

Draw independent SRSs from two populations with population proportions of successes p1p_1 and p2.p_2. To get the **plus four confidence interval for the difference** p1-p2,$p_1 - p_2$, add four imaginary observations: one success and one failure in each of the two samples. Then use the large-sample confidence interval with the new sample sizes (actual sample sizes + 2) and counts of successes (actual counts + 1).

Use this interval when the sample size is at least 5 in each group, with any counts of successes and failures.

If your software does not offer the plus four method, just enter the new plus four sample sizes and success counts into the large-sample procedure.

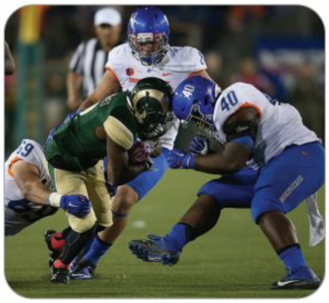
EXAMPLE 20.3 CTE and contact sport

STATE: Chronic traumatic encephalopathy (CTE) is a progressive neurodegenerative disorder that is caused by repetitive brain trauma. Recent evidence has shown that professional football players have an elevated risk of developing CTE. A case-control study examined the brains of deceased men with a diagnosed nongenetic neurodegenerative disorder. The study compared individuals with and individuals without a history of participation in amateur contact sports (such as football, wrestling, and basketball). Of the 66 men who had participated in contact sports, 21 had CTE. None of the 132 control men who had not participated in a contact sport had CTE.[4] Among men with a neurodegenerative disorder, how much more common is CTE in those with a history of playing a contact sport than for those without such a history?

PLAN: Take dead men with a nongenetic neurodegenerative disorder and a documented history of contact sport participation to be Population 1 and dead men with a nongenetic neurodegenerative disorder but no documented history of contact sport participation to be Population 2. The population proportions whose postmortem brains show evidence of CTE are p1p_1 for men with a history of contact sports and p2p_2 for men with no history of contact sports. We want to estimate the difference p1-p2 $p_1 - p_2$ with 95% confidence.

SOLVE: First verify the conditions for inference:

- This case-control observational study uses brains selected from a donated-brain bank in Florida. According to the researchers, the brain bank is fairly representative of the general population of individuals with a neurodegenerative disorder, so we should be safe in considering the data as two independent SRSs from the male portion of this population.
- The conditions for the large-sample interval are not met. In fact, there are *no* "successes" (CTE) in the control group. However, the sample sizes are each greater than 5, so we can safely compute a plus four confidence interval.



Doug Pensinger/Getty Images

Here are the summary statistics for this study:

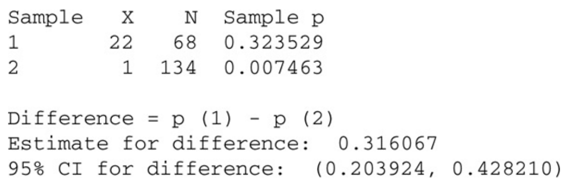
	Population description	Sample size	Number of	Sample proportion

Population			successes	
1	contact sport	n1=66$n_1 = 66$	21	p^1=21/66=0.3182$\hat{p}_1 = 21/66 = 0.3182$
2	no contact sport	n2=132$n_2 = 132$	0	p^2=0/132=0.0000$\hat{p}_2 = 0/132 = 0.0000$

The plus four confidence interval adds four imaginary observations, leading to this new data summary:

Population	Population description	Sample size	Number of successes	Plus fou
1	contact sport	n1+2=68$n_1 + 2 = 68$	21+1=22$21 + 1 = 22$	p1̃=22/68=0
2	no contact sport	n2+2=134$n_2 + 2 = 134$	0+1=1$0 + 1 = 1$	p2̃=1/134=0

With the updated counts, we can use technology to obtain a 95% plus four confidence interval for p1-p2.$p_1 - p_2$. Here is the output from Minitab:



```
Sample    X     N  Sample p
1        22    68  0.323529
2         1   134  0.007463

Difference = p (1) - p (2)
Estimate for difference:  0.316067
95% CI for difference:  (0.203924, 0.428210)
```

CONCLUDE: We are 95% confident that a history of contact sports increases the chance of CTE among men who eventually develop a nongenetic neurodegenerative disease by between 20.4 and 42.8 percentage points. Because the study focused exclusively on men, these findings cannot safely be extended to women.

The computations performed by Minitab in this example involve the following steps (with some roundoff error):

- Sample estimate: p1̃-p2̃=0.32353-0.00746=0.31607
  $\tilde{p}_1 - \tilde{p}_2 = 0.32353 - 0.00746 = 0.31607$
- Standard error of the estimate:

SE=p̃1(1-p̃1)n1+2+p̃2(1-p̃2)n2+2=(0.32353)(0.67647)68+(0.00746)(0.99254)134=0.00327=0.05722

$$\begin{aligned} \mathrm{SE} &= \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2}+\frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}} = \sqrt{\frac{(0.32353)(0.67647)}{68}+\frac{(0.00746)(0.99254)}{134}} \\ &= \sqrt{0.00327} = 0.05722 \end{aligned}$$

- Standard Normal critical value for a 95% confidence level: z*=1.960 $z^* = 1.960$ (rounded to three decimal places, obtained from software or Table C)
- Margin of error: m=z*SE=1.960(0.05722)=0.11215 $m = z^*\mathrm{SE} = 1.960\,(0.05722) = 0.11215$
- 95% confidence interval for p1-p2: (p̃1-p̃2)±m=0.31607±0.11215, $p_1 - p_2 : \; (\tilde{p}_1 - \tilde{p}_2) \pm m = 0.31607 \pm 0.11215,$ or 0.20392 to 0.42822$0.20392 \text{ to } 0.42822$

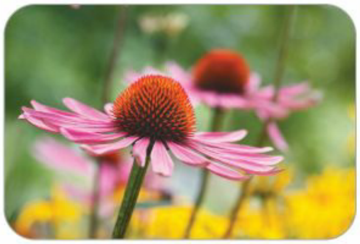
The plus four interval may be conservative (that is, the true confidence level may be *higher* than you asked for) for very small samples and population p'sp's close to 0 or 1, as in this example. This approach is generally much more accurate than the large-sample interval when the samples are small. It is also simple enough that it requires no specialized software. The Minitab function used here is the same that was shown in Figure 20.2 for Example 20.3.

APPLY YOUR KNOWLEDGE

20.3 Echinacea for the common cold? Echinacea is widely used as an herbal remedy for the common cold, but does it work? In a double-blind experiment, healthy volunteers agreed to be exposed to common-cold–causing rhinovirus type 39 and have their symptoms monitored. The volunteers were randomly assigned to take either a placebo or an echinacea supplement daily for 5 days following viral exposure. Among the 103 subjects taking a placebo, 88 developed a cold, whereas 44 of the 48 subjects taking echinacea developed a cold.[5]

a. Explain why the large-sample confidence interval is not appropriate for these data.

b. Give the 95% plus four confidence interval for the difference in proportions of individuals developing a cold after viral exposure between the echinacea treatment and the placebo.



imageBROKER/Alamy Stock Photo

20.4 Vaccine protection against cervical cancer. Most cases of cervical cancer are linked to a few strains of the human papillomavirus (HPV). The pharmaceutical company Merck developed a vaccine (Gardasil) against these HPV strains. Worldwide clinical trials followed young women 16 to 26 years of age after vaccination or administration of a placebo for two to four years for signs of HPV-caused cervical cancer. Of 8460 women given Gardasil, none developed HPV-caused cervical cancer. In contrast, 53 of the 8487 women given a placebo did.[6]

a. Why should we not use the large-sample confidence interval for these data?

b. Give the plus four 95% confidence interval for the difference between the two population proportions of women developing HPV-caused cervical cancer.

HYPOTHESIS TESTS FOR COMPARING PROPORTIONS

An observed difference between two sample proportions can reflect an actual difference between the populations, or it may just be due to chance variation in random sampling. Hypothesis tests help us decide if the effect we see in the samples is really there in the populations. The null hypothesis says that there is no difference between the two population proportions of the outcome of interest (so that p1-p2=0):$p_1 - p_2 = 0)$:

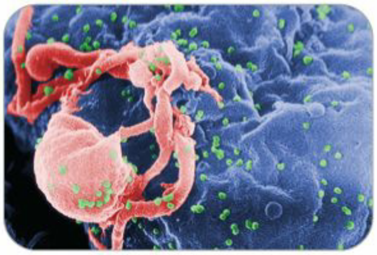
H0: p1=p2

$H_0 : p_1 = p_2$

The alternative hypothesis says which kind of difference we expect.

EXAMPLE 20.4 Toward an HIV vaccine

STATE: After decades of research and clinical trials, a recent study has brought hope that we will one day have an effective vaccination against HIV. The study, conducted in Thailand, enrolled HIV-negative men and women between the ages of 18 and 30 years who were at an average risk of infection. The design was randomized and double-blind, assigning half of the subjects to a placebo and the other half to a combination of two vaccines, ALVAC and AIDSVAX, which on their own had previously not worked. All participants received counseling on how to avoid becoming infected with HIV and were regularly tested for HIV infection over a three-year period. Of 8198 subjects given a placebo, 74 became infected with HIV, compared with 51 of 8197 vaccinated subjects.[7] Do the data support the hypothesis that the vaccination combination is better than a placebo in preventing HIV infections?



Cynthia Goldsmith/CDC

> **PLAN:** Take individuals given a placebo to be Population 1 and vaccinated individuals to be Population 2. Our alternative is clearly one-sided, so we test
>
> H0: p1=p2Ha: p1>p2
>
> $$H_0 : p_1 = p_2$$
> $$H_a : p_1 > p_2$$
>
> **SOLVE:** The study design is randomized, so we can treat the two samples as if they were two independent SRSs of the two populations. The sample proportions who become infected with HIV are
>
> p^1=748198=0.0090 (placebo)p^2=518197=0.0062 (vaccine)
>
> $$\hat{p}_1 = \frac{74}{8198} = 0.0090 \quad \text{(placebo)}$$
> $$\hat{p}_2 = \frac{51}{8197} = 0.0062 \quad \text{(vaccine)}$$
>
> That is, 0.9% of the subjects given a placebo but only 0.6% of the vaccinated subjects became infected with HIV over a three-year period. Is this apparent difference statistically significant? To continue with the solution, we must learn the proper test.

To test the null hypothesis, we standardize p^1-p^2 $\hat{p}_1 - \hat{p}_2$ to get a z z statistic. If H0 H_0 is true, p1=p2 $p_1 = p_2$ and all the observations in both samples come from a single population of Thai adults of whom a single unknown proportion p p would become infected with HIV over a three-year period. Thus, instead of estimating p1 p_1 and p2 p_2 separately, we pool the two samples and use the overall sample proportion to estimate the single population parameter p. p. We call this the **pooled sample proportion**:

p^=number of successes in both samples combinednumber of individuals in both samples combined

$$\hat{p} = \frac{\text{number of successes in both samples combined}}{\text{number of individuals in both samples combined}}$$

Use p^ $\hat{p}$ in place of both p^1 $\hat{p}_1$ and p^2 $\hat{p}_2$ in the expression for the standard error SE of p^1-p^2 $\hat{p}_1 - \hat{p}_2$ to get a z z statistic that has the standard Normal distribution when H0 H_0 is true. As always, the test *P*-value is the probability, if H0 H_0 was true, of obtaining a test statistic at least as extreme (in the direction of Ha H_a) as that actually obtained, and it can be computed as the corresponding area under the standard Normal curve. Here is the whole test.

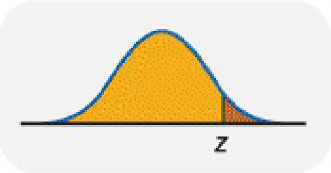
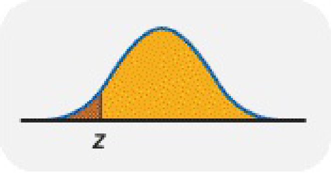
HYPOTHESIS TEST FOR COMPARING TWO PROPORTIONS

Draw an SRS of size n1n_1 from a large population having proportion p1p_1 of successes, and draw an independent SRS of size n2n_2 from another large population having proportion p2p_2 of successes. To **test the hypothesis** H0:p1=p2,$\boldsymbol{H_0 : p_1 = p_2}$, obtain the z$z$ statistic

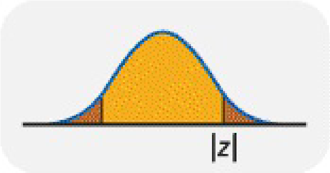
z=p^1-p^2p^(1-p^)1n1+1n2

$$z=\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

where p^$\hat{p}$ is the pooled proportion of successes in both samples combined. In terms of a variable ZZ having the standard Normal distribution, the *P*-value for a test of H0H_0 against

Ha: p1>p2$H_a : p_1 > p_2$	isis	P(Z≥z)$P(Z \geq z)$	 Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company
Ha: p1<p2$H_a : p_1 < p_2$	isis	P(Z≤z)$P(Z \leq z)$	 Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Ha: p1≠p2$H_a : p_1 \neq p_2$ isis 2P(Z≥|z|)$2P(Z \geq |z|)$



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

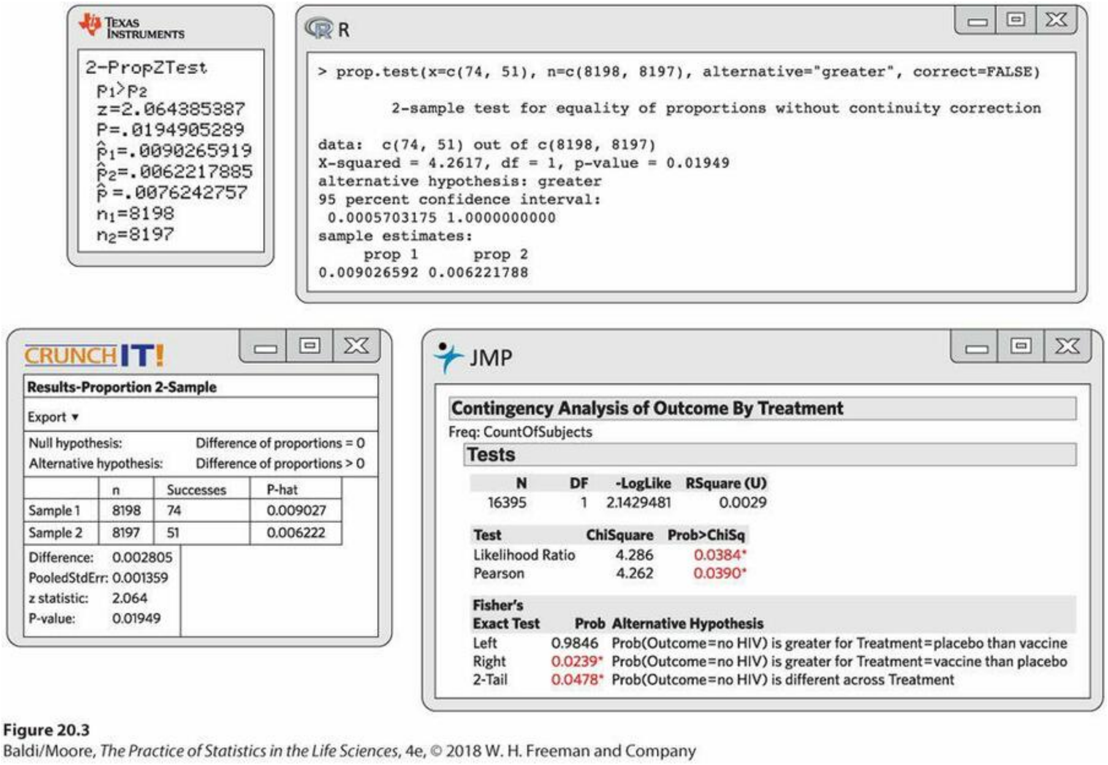
Technology with basic statistical inference capabilities will perform all necessary computations and provide both the test statistic and the *P*-value.

Use this test when the counts of successes and failures are each 5 or more in both samples.[8]

EXAMPLE 20.5 Toward an HIV vaccine, continued

SOLVE: The data come from a randomized experiment, and the counts of successes and failures are all larger than 5. We can safely run a two-sample zz test for proportions.

We see from the TI-83 graphing calculator and CrunchIt! software output in Figure 20.3 that the test statistic is z=2.064$z = 2.064$ and the one-sided *P*-value is 0.01949, a value typically considered statistically significant.



TEXAS INSTRUMENTS

```
2-PropZTest
 p1>p2
 z=2.064385387
 P=.0194905289
 p̂1=.0090265919
 p̂2=.0062217885
 p̂ =.0076242757
 n1=8198
 n2=8197
```

R

```
> prop.test(x=c(74, 51), n=c(8198, 8197), alternative="greater", correct=FALSE)

        2-sample test for equality of proportions without continuity correction

data:  c(74, 51) out of c(8198, 8197)
X-squared = 4.2617, df = 1, p-value = 0.01949
alternative hypothesis: greater
95 percent confidence interval:
 0.0005703175 1.0000000000
sample estimates:
     prop 1      prop 2
0.009026592 0.006221788
```

CRUNCH IT!

```
Results-Proportion 2-Sample
Export ▾
Null hypothesis:          Difference of proportions = 0
Alternative hypothesis:   Difference of proportions > 0

           n      Successes    P-hat
Sample 1   8198   74           0.009027
Sample 2   8197   51           0.006222
Difference:    0.002805
PooledStdErr:  0.001359
z statistic:   2.064
P-value:       0.01949
```

JMP

```
Contingency Analysis of Outcome By Treatment
Freq: CountOfSubjects
  Tests
        N      DF   -LogLike   RSquare (U)
    16395       1   2.1429481       0.0029

  Test              ChiSquare   Prob>ChiSq
  Likelihood Ratio      4.286      0.0384*
  Pearson               4.262      0.0390*

  Fisher's
  Exact Test      Prob   Alternative Hypothesis
  Left          0.9846   Prob(Outcome=no HIV) is greater for Treatment=placebo than vaccine
  Right         0.0239*  Prob(Outcome=no HIV) is greater for Treatment=vaccine than placebo
  2-Tail        0.0478*  Prob(Outcome=no HIV) is different across Treatment
```

Figure 20.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 20.3 Output from a graphing calculator and three statistical programs for a study of the efficacy of a new HIV vaccine.

CONCLUDE: There is good evidence (P≈0.02)$(P \approx 0.02)$ that the HIV infection rate is significantly lower among vaccinated individuals than among individuals given a placebo. The HIV vaccine combo offers better protection against HIV infection than a placebo in this population.

Here are the computations required to obtain the test *P*-value:

- Pooled proportion of individuals getting infected with HIV:

p^=number of individuals infected with HIV in both samplesnumber of individuals in both samples combined=74+518198+8197=12516,395=0.0076

$$\begin{aligned}\hat{p} &= \frac{\text{number of individuals infected with HIV in both samples}}{\text{number of individuals in both samples combined}}\\ &= \frac{74+51}{8198+8197}=\frac{125}{16{,}395}=0.0076\end{aligned}$$

- Test statistic:

z=p^1-p^2p^(1-p^)1n1+1n2=0.0090-0.0062(0.0076)(0.9924)18198+18197=0.00280.00136=2.06

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.0090-0.0062}{\sqrt{(0.0076)(0.9924)\left(\frac{1}{8198}+\frac{1}{8197}\right)}}$$
$$= \frac{0.0028}{0.00136} = 2.06$$

- *P*-value: The one-sided *P*-value for this test is the upper-tail area under the standard Normal curve for values of z≥2.064, $z \geq 2.064$, as illustrated in Figure 20.4. This area can be computed with technology using commands such as normalcdf(2.064,1E99,0,1) in the TI-83/84 calculator or =1-NORM.S.DIST(2.064,TRUE) in Excel, which give a *P*-value of 0.0195. The z* z^* row of Table C tells us only that the one-sided *P*-value is between 0.01 and 0.02 because z=2.064 $z = 2.064$ is between the critical *z*-values 2.054 and 2.326.

z* z^*	2.054	2.326
One-sided *P*	.02	.01

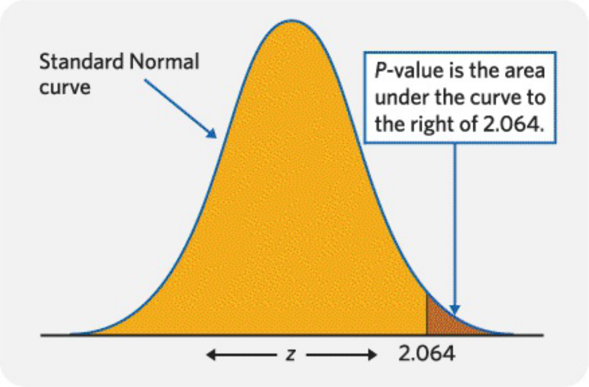


Figure 20.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 20.4 The *P*-value for the one-sided test for a study of the efficacy of a new HIV vaccine.

CrunchIt! and the TI-83 graphing calculator both use the Normal approximation for the sampling distribution of p^1-p^2. $\hat{p}_1 - \hat{p}_2$. Notice, however, that R and JMP do not provide the value of the test statistic z. z. Instead, they report a value “X-squared=4.2617” $\text{“X-squared} = 4.2617\text{”}$ (or “Pearson ChiSquare 4.262” for JMP), which is the squared value of z=2.064385. $z = 2.064385$. This method is called the chi-square test, and we will study it further in Chapter 22. When the categorical variable is binary (“success” or “failure”), the chi-square test is the equivalent of the zz test for two proportions with a two-sided alternative. JMP reports a generic, two-sided *P*-value (0.0390=2×0.0195),$(0.0390 = 2 \times 0.0195)$, whereas R provides the one-sided *P*-value we want since we specified “alternative="greater".” For any test, remember to always specify your null and alternative hypotheses before examining the sample data and then obtain the one- or two-sided *P*-value that matches your alternative.

APPLY YOUR KNOWLEDGE

20.5 Neonatal care in Malawi. There are safe and effective interventions to treat respiratory failure in newborns, but they are typically costly. The lack of widespread access to these costly interventions is one reason that respiratory failure is still a leading cause of neonatal death in developing countries. A new and cheaper bubble continuous positive airway pressure (bCPAP) system could be a more affordable option. In a low-resource area in Malawi, researchers compared the survival of infants with severe respiratory distress who were treated, at random, either with the new bCPAP machine or with nasal oxygen therapy (the current local standard of care, serving as a control).[9] The published findings report that “survival rate for neonates receiving bCPAP was 71.0% (44/62) compared with 44.0% (11/25) for controls.”

a. Define the individuals studied, the type of study, and the explanatory and response variables. Determine whether the variables are quantitative or categorical.

b. The researchers tested “the hypothesis that the rate of survival was higher for babies receiving bCPAP than in the control group.” State the corresponding null and alternative hypotheses using mathematical symbols.

c. Check whether the conditions for inference are met. Note that, for ethical reasons, the researchers provided care with the bCPAP machine to all newborns who needed it unless the machine was not available. This is not quite a true random assignment but it is unlikely to be biased, so we may consider it to be approximately random.

d. Obtain the test statistic and *P*-value, then conclude in context.

20.6 FDA warning on insomnia prescriptions. In a January 2013 safety announcement, the U.S. Food and Drug Administration (FDA) warned drug manufacturers that the recommended dose of the widely prescribed insomnia drug zolpidem should be halved for women. The warning came after pharmacokinetic trials found that 15% of 250 women and 3% of 250 men who had taken the recommended dose of zolpidem still had enough drug in their blood 8 hours later to cause impairment. Do the data support the claim that women are more likely than men to have impairing zolpidem levels 8 hours after dosing? Follow the four-step process as illustrated in Examples 20.4 and 20.5.

AMBIEN CR
ZOLPIDEM TARTRATE EXTENDED-RELEASE
12.5 mg Tablets
Dispense with Medication Guide
Rx only
100 Tablets
sanofi aventis

JB Reed/Bloomberg via Getty Images

RELATIVE RISK AND ODDS RATIO[i]

Categorical variables are particularly widespread in the health sciences. Many studies describe the presence or absence of a disease in a particular population or an improvement or a lack of improvement after treatment. In fact, the objective of most scientific reports in this field is to study either risk factors or treatment efficacy.

Studies evaluating the potential impact of a risk factor look for an association between exposure to the risk factor and a negative health outcome in sample individuals. As discussed in Chapter 6, these epidemiological studies can have different designs, but they are typically observational. Example 20.3 describes one such case-control observational study of the association between CTE and a history of participating in contact sports.

Studies evaluating the efficacy of a treatment compare the outcome of a sample given the treatment with the outcome of a sample given a placebo or some other control. These are typically experimental studies, whose design was covered in detail in Chapter 7. Example 20.5 is one such randomized experiment comparing the effectiveness of a new HIV vaccine and a placebo.

Both types of medical studies examining a categorical variable in two populations rarely report the difference in sample proportions as we have studied in this chapter so far. Instead, you typically read about a “relative risk” or an “odds ratio” computed from the sample data. The following section gives a brief introduction to these topics.

Relative risk and odds ratio Both observational and experimental studies in the health sciences typically compare a sample from a population of interest with a sample from a control population. The comparison is then represented by either a relative risk or an odds ratio.

Chapter 9 introduced the concepts of risk and odds in probability. When comparing two groups, the relative risk, RR,$RR,$ is simply the ratio of the two risks with the control group in the denominator. Likewise, the odds ratio, OR,$OR,$ is the ratio of both odds with the control group in the denominator. For example, if a study finds that 50% of individuals in the treatment group had some symptoms compared with only 25% in the control group, the relative risk would be RR=50%/25%=2$RR = 50\%/25\% = 2$ and the odds ratio would be OR=(1/1)/(1/3)=3.$OR = (1/1)/(1/3) = 3.$ Relative risk is easy to interpret—individuals in the treatment group are twice as likely to have symptoms as similar control individuals. Interpreting the odds ratio is more challenging, because it compares unfamiliar odds rather than probabilities. Odds ratios can also give an exaggerated impression of the relative difference between the groups, particularly when the events studied are not rare. However, odds ratios offer mathematical advantages for advanced statistical analysis. Odds ratios are found most often in observational, case-control (or historical control) epidemiological studies, while relative risks are typically used in prospective cohort studies and randomized controlled experiments.

RELATIVE RISK AND ODDS RATIO IN THE HEALTH SCIENCES

Draw an SRS of size n1n_1 from a large population having proportion p1p_1 of a medical outcome, and draw an independent SRS of size n2n_2 from another large population having proportion p2p_2 of that outcome. Let the first sample represent the group of interest and the second sample represent the reference, or control, group.

The relative risk for the medical outcome in the group of interest compared with the reference group is

RR=p^1p^2

$$RR = \frac{\hat{p}_1}{\hat{p}_2}$$

The relative risk is sometimes also called the hazard ratio, or HR.$HR.$

The odds ratio for the medical outcome in the group of interest compared with the reference group is

OR=odds1odds2=p^1(1-p^2)p^2(1-p^1)

$$OR = \frac{odds_1}{odds_2} = \frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)}$$

EXAMPLE 20.6 Preventing blood clots in immobilized patients

Patients who remain immobilized for a substantial amount of time can develop deep vein thrombosis (DVT), a blood clot in a leg or pelvis vein. DVT can have serious adverse health effects and can be difficult to diagnose. On its website, the drug manufacturer Pfizer reports the outcome of a study looking at the effectiveness of the drug Fragmin (dalteparin) compared with that of a placebo in preventing DVT in immobilized patients.

In a double-blind, multinational study, severely immobilized patients were randomly assigned to receive daily subcutaneous injections of either Fragmin or a placebo for 12 to 14 days and were monitored for 90 days. The results, in number of patients experiencing a complication from DVT (including death), are summarized in the following table:

15
20
10.3
4.0/1
3:27
SP
SL
Fo
1
Co

James King-Holmes/Science Source

	Outcome		
Treatment	**Complication**	**No complication**	**Sample size**
Fragmin	42	1476	1518
Placebo	73	1400	1473

The proportion of subjects experiencing DVT complications in the two samples are

p^Fragmin=42/1518=0.0277p^placebo=73/1473=0.0496

$$\begin{aligned}\hat{p}_{Fragmin} &= 42/1518 = 0.0277\\ \hat{p}_{placebo} &= 73/1473 = 0.0496\end{aligned}$$

The relative risk of DVT complications is the ratio of the two sample proportions, with the placebo group in the denominator:

RR=p^Fragminp^placebo=0.02770.0496=0.558

$$RR = \frac{\hat{p}_{Fragmin}}{\hat{p}_{placebo}} = \frac{0.0277}{0.0496} = 0.558$$

That is, patients receiving the Fragmin injections are less likely to develop complications from DVT than patients in the control group.

The odds of a subject experiencing DVT complications in the two samples are

oddsFragmin=42/1476,or about 1 to 35 againstoddsplacebo=73/1400,or about 1 to 19 against

$$\begin{aligned} odds_{Fragmin} &= 42/1476, \text{or about 1 to 35 against} \\ odds_{placebo} &= 73/1400, \text{or about 1 to 19 against} \end{aligned}$$

The odds ratio of DVT complications is the ratio of these two odds, with the placebo group in the denominator, but it can also be computed from the two sample proportions:

OR=oddsFragminoddsplacebo=42/147673/1400=0.546=p^Fragmin(1-p^placebo)p^placebo(1-p^Fragmin)=0.0277×(1-0.0496)0.0496×(1-0.0277)=0.546

$$\begin{aligned} OR &= \frac{odds_{Fragmin}}{odds_{placebo}} = \frac{42/1476}{73/1400} = 0.546 \\ &= \frac{\hat{p}_{Fragmin}(1-\hat{p}_{placebo})}{\hat{p}_{placebo}(1-\hat{p}_{Fragmin})} = \frac{0.0277\times(1-0.0496)}{0.0496\times(1-0.0277)} = 0.546 \end{aligned}$$

This value is quite similar to the RR$RR$ value. When samples are large and the outcome studied is rare (here 3% to 5%), RR$RR$ and OR$OR$ are very close.

What do the values for an RR$RR$ or an OR$OR$ mean? A relative risk of 1 means that there is no difference in risk between the two groups. However, relative risk is computed from sample data and is only an estimate of the true population risks. Therefore, an RR$RR$ different from 1 does not necessarily imply a difference between the two populations. How large or how small a relative risk—or an odds ratio—needs to be so as to be statistically significant depends on the sampling distributions of these statistics.

Because relative risk and odds ratio are expressed as ratios of random variables, their sampling distributions are not symmetric. Therefore, confidence intervals for the population parameter are not symmetric and cannot be expressed as “estimate ±$\pm$ *margin of error.”* For instance, in the Fragmin study described in Example 20.6, special software intended for the analysis of medical data gives the following confidence intervals:

```
Statistic        Value     Low 95% CI   High 95% CI
Odds ratio       0.546       0.364         0.816
Relative risk    0.558       0.378         0.823
```

The calculations required to obtain such confidence intervals are beyond the scope of this introductory statistics textbook. The interpretation of a confidence interval or a *P*-value is the same, however, whether we study an average, a proportion, an odds ratio, or a relative risk. As always, *statistical significance does not imply relevance or importance. When sample sizes are very large, statistical significance can be reached for even very small effects. Moreover, observational studies always carry potential confounding factors that make a clear-cut interpretation of cause and effect between a studied risk factor and an observed outcome impossible.*

In the Fragmin experiment, because the 95% confidence interval for RR$RR$ contains only values less than 1, we can conclude with 95% certainty that severely immobilized patients given Fragmin are less likely (lower risk) to experience DVT complications than similar patients given a placebo.

APPLY YOUR KNOWLEDGE

20.7 Aspirin and heart attacks. The landmark Physicians’ Health Study randomly assigned 22,071 healthy male physicians at least 40 years old to take either an aspirin every other day or a placebo pill every other day. Of the 11,037 physicians who took aspirin regularly for 5 years, 10 had a fatal heart attack. Of the 11,034 who took the placebo, 26 suffered a fatal heart attack.[10]

a. What are the proportions of physicians who suffered a fatal heart attack in the two groups?

b. Calculate the relative risk of a fatal heart attack for physicians taking aspirin compared with those taking a placebo. Explain what this value means in context.

c. Calculate the odds ratio for a fatal heart attack comparing physicians taking aspirin to those taking a placebo. How does this value compare with the RR$RR$ you calculated in part b?

20.8 Antibiotics in the ICU. Patients who are admitted to the intensive care unit (ICU) and require mechanical ventilation can develop fatal bacterial infections. A clinical trial randomly assigned 934 patients admitted to an ICU and requiring mechanical ventilation to two groups. The control group received a standard medical treatment, whereas the antibiotics group received antibiotics prophylactically in addition to the standard treatment. The study found that 107 of the 468 patients in the control group died, compared with 69 of the 466 patients in the antibiotics group.[11]

a. What are the proportions of ICU patients who died in the two groups?

b. Calculate the relative risk of death for ICU patients given the antibiotics prophylactically compared with those given only the standard treatment. Explain what this value means in context.

c. Calculate the odds ratio for death, comparing ICU patients given the antibiotics prophylactically compared with those given only the standard treatment. How does this value compare with the RR$RR$ you calculated in part b?

Effect size: relative versus absolute measures. The relative risk gives a measure of how effective a treatment is or how influential a risk factor is in a group compared with a control group without the treatment or the risk factor. One topic of debate in the biomedical community centers on how to interpret this value. We will compare the relative risk to other measures of treatment effectiveness in comparative, randomized experiments.

RELATIVE RISK REDUCTION, ABSOLUTE RISK REDUCTION, AND NUMBER NEEDED TO TREAT

In a randomized clinical trial comparing the proportion of a negative medical outcome in a group provided a medical treatment and in a control group:

The relative risk reduction, RRR,RRR, for the treatment group compared with the control group is the reduction in risk expressed as a fraction of the control group's risk:

RRR=p^control-p^treatmentp^control

$$RRR = \frac{\hat{p}_{control} - \hat{p}_{treatment}}{\hat{p}_{control}}$$

The absolute risk reduction, ARR,ARR, for the treatment group compared with the control group is simply the difference in risks:

ARR=p^control-p^treatment

$$ARR = \hat{p}_{control} - \hat{p}_{treatment}$$

The number needed to treat, NNT,NNT, is the expected number of subjects who must be treated before 1 subject in the treatment group can be spared the negative medical outcome, compared with what would be expected in the control group. When ARR$ARR$ is a strictly positive value, NNT$NNT$ is the inverse of ARR:ARR:

NNT=1ARR=1p^control-p^treatment

$$NNT = \frac{1}{ARR} = \frac{1}{\hat{p}_{control} - \hat{p}_{treatment}}$$

(When studying treatment side effects, ARR$ARR$ is typically negative and the number needed to harm, NNH,NNH, is reported instead.)

EXAMPLE 20.7 Preventing blood clots: relative risk reduction

Example 20.6 described the efficacy of Fragmin, compared with a placebo, for preventing complications from DVT in immobilized patients. Here are the proportions of adverse events in the two groups:

p^Fragmin=42/1518=0.0277p^placebo=73/1473=0.0496

$$\begin{aligned}\hat{p}_{Fragmin} &= 42/1518 = 0.0277\\ \hat{p}_{placebo} &= 73/1473 = 0.0496\end{aligned}$$

The Fragmin website reports a relative risk reduction of 44%. How was this number computed?

RRR=p^placebo-p^Fragminp^placebo=0.0496-0.02770.0496=0.442

$$RRR = \frac{\hat{p}_{placebo} - \hat{p}_{Fragmin}}{\hat{p}_{placebo}} = \frac{0.0496 - 0.0277}{0.0496} = 0.442$$

That is, immobilized patients treated with Fragmin have a 44% lower probability of suffering complications from DVT than similar patients given a placebo instead.

How does this value compare with the absolute risk reduction for this study?

ARR=p^placebo-p^Fragmin=0.0496-0.0277=0.0219

$$ARR = \hat{p}_{placebo} - \hat{p}_{Fragmin} = 0.0496 - 0.0277 = 0.0219$$

Compared with receiving a placebo, receiving Fragmin injections helps prevent DVT complications in roughly 2% of all immobilized patients.

The number needed to treat, NNT,NNT, is gradually becoming a standard in communicating clinical results. In the Fragmin study, this value is

NNT=1ARR=10.0219=45.7

$$NNT = \frac{1}{ARR} = \frac{1}{0.0219} = 45.7$$

That is, we expect that, on average, for every 46 immobilized patients treated with Fragmin instead of a placebo, DVT complications can be prevented in 1 patient.

The 2% ARR$ARR$ is radically different from the 44% RRR,RRR, and both give quite a different impression of treatment success. RRR$RRR$ tends to inflate the impression of success, but ARR$ARR$ can be difficult to comprehend. There is a growing trend toward always providing the NNT NNT in medical studies, as it gives a clear, practical interpretation of the relative benefit of treatment. Of course, the decision to provide a preventive or curative treatment depends not only on the relative benefit of one particular treatment, but also on such issues as the availability, costs, and benefits of alternative treatments, the number and seriousness of potential adverse events, and the seriousness and consequences of the condition (for example, death versus minor discomfort).

APPLY YOUR KNOWLEDGE

20.9 Aspirin and heart attacks, continued. Go back to Exercise 20.7 about the Physicians' Health Study.

a. Calculate the absolute and relative risk reduction in fatal heart attacks for physicians taking aspirin compared with those taking a placebo. Explain briefly what these two values represent.
b. Calculate the number needed to treat with aspirin compared with a placebo to prevent 1 fatal heart attack. What does this value tell you?

20.10 Antibiotics in the ICU, continued. Go back to Exercise 20.8 about the prophylactic use of antibiotics in the ICU.

a. Calculate the absolute and relative risk reduction in death for ICU patients given the antibiotics prophylactically compared with those in the control group. Explain briefly what these two values represent.
b. Calculate the number needed to treat with a prophylactic antibiotic regimen compared with doing nothing beyond a standard treatment to prevent 1 death. What does this value tell you?

DISCUSSION Assessing and understanding health risks

We have all heard news stories about factors that may increase or decrease the risk of getting a particular disease. At times, press releases even seem conflicting. For instance, we hear that coffee may reduce the risk of heart disease but also that it may increase the risk of a heart attack. So where does this information come from? And what does it really mean? Epidemiological studies and clinical trials are two major tools used to study health risk. Here we discuss how health risks are assessed in these two different contexts and what information they really convey.

Studying risk factors in epidemiological studies

Epidemiology is the science of understanding disease patterns through observations. Epidemiologists examine data to see what behaviors, locations, or exposures are associated with various diseases and causes of death. Because these studies are observational, confounding is a common problem and causality cannot be directly established. As discussed in Chapter 7, there are several types of epidemiological studies. Prospective cohort studies are the least prone to confounding errors, because subject selection does not depend on the factors studied. Despite their weaknesses, epidemiological studies can be very effective when many different studies converge to the same conclusion, as with tobacco smoking and the risk of developing lung and several other cancers.

On the one hand, because confounding can never be completely ruled out in observational studies, finding a significant difference between some groups does not necessarily imply that a true risk factor has been discovered. This is especially true if the effect found is small. For example, a Harvard study found a two-and-a-half times higher risk of pancreatic cancer among coffee drinkers than among non-coffee drinkers. However, the association was not substantiated by later research from this group and countless others. By comparison, smoking multiplies the risk of cancer 20 to 30 times. On the other hand, even a small increase that shows up consistently in many studies with different designs can be a credible finding and may have a large societal impact if the risk is widespread. Meta-analysis is an advanced statistical technique developed to find significant trends in the results of many different studies of the same variable. It has become a fairly common and important tool of epidemiology.

The information about new health risks that reaches the general public depends on scientific publications and how they are reported in the media. Unfortunately, studies that find no evidence of an association are often unreported or altogether abandoned. Significant results can be published more easily and in more prestigious papers and are much more likely to be funded and reported in the news. This is an important bias that cannot be offset by running a meta-analysis. A related and potentially devastating bias is due to financial conflicts of interest. Studies show that results may go unreported or conclusions may be unreasonably speculative when such conflicts are present. For example, a 1998 publication reporting on 12 patients speculated about a possible link between the measles, mumps, rubella (MMR) vaccine and autism, creating quite a stir.[12] It was later revealed that the lead author had received funding from lawyers who had filed class action suits against the vaccine manufacturer and that most of the “findings” in the published report had been faked.[13] Many well-designed studies carried out since then have found no evidence of a link between the MMR vaccine and autism, yet public skepticism persists and has prompted many parents worldwide to refuse to vaccinate their children. The public can be scared easily by worrisome reports, and it may take a whole generation to undo the damage.

STATS IN YOUR WORLD

SOUNDS GOOD–BUT NO COMPARISON

Most women have mammograms to check for breast cancer once they reach middle age. Could a fancier test do a better job of finding cancers early? Positron emission tomography (PET) scans are a fancier (and more expensive) test. Doctors used PET scans on 14 women with tumors and got the detailed diagnosis right in 12 cases. That’s promising. But there were no controls, and 14 cases are not statistically significant. Medical standards require randomized comparative experiments and statistically significant results. Only then can we be confident that the fancy test really is better.

Health risks and treatment options in clinical trials

Clinical trials are randomized controlled experiments involving human subjects. For obvious ethical reasons, clinical trials are limited to studying ways to reduce risk and not risk factors themselves. Because subjects are randomly assigned to various treatments and controls, differences in response can be directly attributed to differences in treatment. However, study participants are typically volunteers, and they may not always represent the target population adequately. The undercoverage of women and minority groups in clinical trials, for instance, is a serious issue, because trial results are often generalized to the whole target population despite their poor representation.

Today patients have become increasingly involved in making their own treatment decisions, especially since the appearance of internet health sites and pharmaceutical advertising. So how do patients make sense of claims about risk reduction for a given treatment? Studies show that our understanding of risk is strongly influenced by the way it is presented and framed. Like physicians and funding institutions, patients are much more likely to favor treatment when the information about risk is presented in a relative way—for instance, by using relative risk reduction, RRR.RRR. Patients also tend to choose treatment more often when risk is described in terms of negative outcomes (death or disease) rather than as a positive outcome (survival or remaining event free).

To consider these issues in context, let’s examine a clinical trial for pravastatin, a cholesterol-lowering drug and one of the best-selling prescription drugs in the United States (under its brand names). The study can be summarized with the statement that (a) pravastatin leads to a 22% reduction of the risk of coronary death (RRR),(RRR), (b) the reduction is 0.9 percentage points (absolute risk reduction, ARR$ARR$), or (c) 111 patients need to take pravastatin daily for five years to prevent 1 coronary death (number needed to treat, NNT$NNT$). These three statements focus on the negative outcome, coronary death. The information can also be reframed to say that taking the drug daily for five years increases the probability of *avoiding* coronary death from 95.9% to 96.8%. All four summaries reflect the same data, but they give very different impressions. Historically, the relative risk reduction has been predominant, largely pushed by the pharmaceutical industry. However, there is a growing recognition of the need to provide a more comprehensive and balanced description of health risks.

One reason why the way in which risk is presented can be so influential is that patients have difficulty grasping the difference between average risk and personal outcomes. They may perceive a stated percent risk reduction as a guaranteed benefit. In reality, there is no such guarantee. In fact, only a fraction of all patients treated will avoid developing the disease thanks to the treatment; all other patients would have the same outcome (disease or not) regardless of whether they took the treatment. There is no way to tell what will happen to any given patient. At the same time, some proportion of all patients treated will experience negative side effects because of the treatment.

Patients should also understand that clinical trials are conducted over a defined time period and that this limitation impacts risk calculations. For instance, a woman’s risk of breast cancer over the next 10 years is substantially lower than her risk of breast cancer over a lifetime, yet this distinction is rarely emphasized. Long-term adverse events due to treatment are often monitored more extensively in phase IV clinical trials, after a treatment has been made publicly available. Thus, a dilemma exists between bringing the benefits of a new treatment to the target population as soon as possible and remaining cautious about side effects that may be too rare to see in limited trials or may take many years to develop. The withdrawal of the pain reliever Vioxx and warnings about suicide attempts among teenagers taking antidepressants are unfortunate examples of side effects that were discovered much too late.

Health claims and you

The take-home message is that being properly informed is critical to understanding health risks. Clearing up misconceptions about risk and risk reduction should be an important part of the patient–physician dialogue, but this step requires educating physicians about the meaning of risk and how to clearly communicate it. If the information matters to you, you should find the original publication and assess the study design and its possible flaws. No single study is entirely conclusive by itself, so you may also want to examine the findings from related studies.

You should also recognize that over-the-counter supplements and herbal remedies are available for sale without ever having to undergo the rigors of a clinical trial to establish their effectiveness or safety. That is, any promise of health benefit or health risk reduction does not have to be substantiated scientifically (instead, advertisement tends to rely heavily on hand-picked testimonies). Also, because they are sold without a prescription, they come without medical or pharmaceutical counsel. You can find some objective information on the National Center for Complementary and Alternative Medicine website (nccam.nih.gov). A few brands do carry a quality seal, but this label typically represents only a voluntary evaluation of manufacturing standards (for example, to attest for a lack of dangerous contaminants). When it comes to supplements, the recommendation is a simple one: Buyers beware!

CHAPTER 20 SUMMARY

- The data in a **two-sample problem** are two independent SRSs, each drawn from a separate population.
- Tests and confidence intervals to compare the proportions p1p_1 and p2p_2 of successes in the two populations are based on the difference p^1-p^2 $\hat{p}_1 - \hat{p}_2$ between the sample proportions of successes in the two SRSs.
- When the sample sizes n1n_1 and n2n_2 are large, the sampling distribution of p^1-p^2$\hat{p}_1 - \hat{p}_2$ is close to Normal, with mean p1-p2. $p_1 - p_2$.
- Technology can be used to automate the Normal computations required to obtain a confidence interval for p1-p2$p_1 - p_2$ or to test H0: p1=p2. $H_0 : p_1 = p_2$.
- A level CC **large-sample confidence interval for** p1-p2$p_1 - p_2$ is

  (p^1-p^2)±z*SE

  $$(\hat{p}_1 - \hat{p}_2) \pm z^* \text{SE}$$

  where the standard error of p^1-p^2$\hat{p}_1 - \hat{p}_2$ is

  SE=p^1(1-p^1)n1+p^2(1-p^2)n2

  $$\text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  and z*z^* is a standard Normal critical value.
- The true confidence level of a large-sample interval can be substantially less than the planned level C.C. Use this interval only if the counts of successes and failures in both samples are each 10 or greater.
- To get a more accurate confidence interval, add four imaginary observations: one success and one failure in each sample. Then use the same formula or technology to obtain a **plus four confidence interval**. You can use this approach whenever the samples have 5 or more observations each.

- A **hypothesis test for** H0:p1=p2$\boldsymbol{H_0 : p_1 = p_2}$ is based on the value of the zz statistic z=p^1-p^2p^(1-p^)1n1+1n2$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

  where p^$\hat{p}$ is the **pooled sample proportion**

  p^=number of successes in both samples combinednumber of individuals in both samples combined

  $$\hat{p} = \frac{\text{number of successes in both samples combined}}{\text{number of individuals in both samples combined}}$$

  The test *P*-value is the area under the standard Normal distribution corresponding to the probability, when H0H_0 is true, of getting a test statistic zz at least as extreme (in the direction of HaH_a) as that computed. Use this test when there are 5 or more successes and 5 or more failures in each sample.

- In the health sciences, the value p^2-p^1$\hat{p}_2 - \hat{p}_1$ is called the **absolute risk reduction** (ARR)(ARR) for a medical outcome in a target population (1) compared with a reference population (2). Other measures for comparing proportions are also used in this context:
  - The **relative risk** is

    RR=p^1p^2

    $$RR = \frac{\hat{p}_1}{\hat{p}_2}$$

  - The **odds ratio** is

    OR=p^1(1-p^2)p^2(1-p^1)

    $$OR = \frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)}$$

  - The **relative risk reduction** is

    RRR=p^2-p^1p^2

    $$RRR = \frac{\hat{p}_2 - \hat{p}_1}{\hat{p}_2}$$

  - The **number needed to treat** is

    NNT=1ARR=1p^2-p^1

$$NNT = \frac{1}{ARR} = \frac{1}{\hat{p}_2 - \hat{p}_1}$$

- Odds ratios are found most often in observational, retrospective studies, whereas RR,RR,RRR,RRR, and NNT$NNT$ are typically used in prospective designs such as randomized controlled experiments. The number needed to treat is the number of subjects that must be treated before 1 subject in the treatment group can be spared a negative medical outcome compared with what would have been expected in the control group.

THIS CHAPTER IN CONTEXT

In Chapter 19 we computed confidence intervals and tested hypotheses about the proportion p of successes in one population. This chapter expands the concept and z procedures to situations where we want to compare the proportions of successes p_1 and p_2 in two distinct populations. The computations described in this chapter build on the Normal calculations you learned in Chapter 11, provided that the counts of successes and failures in the samples are large enough.

When comparing two populations using two independent random samples, it is important to keep in mind that what you can conclude depends on how the data were collected. No matter how statistically significant a *P*-value is, you cannot directly conclude about causation if the data came from an observational study, because of confounding variables (as emphasized in Chapters 6 and 7). Statistical significance also doesn’t imply practical significance, and it is important to note the effect size as well as its statistical significance. This chapter briefly discusses approaches that help evaluate the relative and absolute differences between two population proportions (RR, OR, RRR, NNT), an important first step in assessing practical relevance.

In Chapters 21 and 22 we will describe another method (chi-square) to test hypotheses for categorical variables beyond the simple case of binary success/failure. That is, we will look at categorical variables that can have more than two levels, such as a person’s blood type (A, B, AB, O) or weight status (underweight, normal weight, overweight, obese). You will notice the parallels between the one-sample z test of Chapter 19 and the chi-square test for goodness of fit of Chapter 21, and between the two-sample z test described in this chapter and the chi-square test for two-way tables of Chapter 22.

CHECK YOUR SKILLS

The culinary herb cilantro, Coriandrum sativum, *is very polarizing: Some people love it and others hate it. A genetic component is suspected in this phenomenon. A survey of American adults of European ancestry found that 1129 of 7295 women and 865 of 7309 men said that cilantro tastes soapy.*[14] *Exercises 20.11 to 20.15 are based on this survey.*

20.11 Take p_W and p_M to be, respectively, the proportions of all women and men among American adults of European ancestry who think that cilantro tastes soapy. We want to know if women and men in this population differ in how they perceive the taste of cilantro. What are the corresponding hypotheses?

a. $H_0: p_W = p_M$ versus $H_a: p_W \neq p_M$
b. $H_0: p_W = p_M$ versus $H_a: p_W > p_M$
c. $H_0: p_W = p_M$ versus $H_a: p_W < p_M$

20.12 What is the pooled sample proportion of women and men who say that cilantro tastes soapy?

a. $\hat{p} = 0.50$
b. $\hat{p} = 0.27$
c. $\hat{p} = 0.14$

20.13 What is the z statistic(in absolute value) for a test comparing the proportions of women and men in this population who say that cilantro tastes soapy?

a. $z = 6.41$
b. $z = 3.20$
c. $z = 1.48$

20.14 The *P*-value for this test is

a. greater than 0.05.
b. between 0.005 and 0.05.
c. less than 0.005.

20.15 What is the margin of error of the 95% large-sample confidence interval for the difference pW-pM$p_W - p_M$ in the proportions of women and men in this population who say that cilantro tastes soapy?

a. 0.048

b. 0.022

c. 0.011

In an experiment to determine whether Substance M can help restore memory, the brains of 20 rats were treated to damage their memories. The rats were then trained to run a maze. After a day, 10 rats were given Substance M and 7 of them succeeded in the maze; only 2 of the 10 control rats were successful. Exercises 20.16 to 20.18 are based on this study.

20.16 The zz test for "no difference" in this case

a. may be inaccurate because the populations are too small.

b. may be inaccurate because some counts of successes and failures are too small.

c. is reasonably accurate because the conditions for inference are met.

20.17 What is the value at the center of the plus four 90% confidence interval for the difference between the proportion of rats that succeed when given Substance M and the proportion that succeed without it?

a. 0.455

b. 0.417

c. 0.5

20.18 What is the margin of error of the plus four 90% confidence interval described in Exercise 20.17?

a. 0.312

b. 0.304

c. 0.185

Glycoprotein IIb/IIIa inhibitors are platelet aggregation inhibitors used during angioplasty for the treatment of heart attacks. You read that eptifibatide is an effective and less expensive GPIIb/IIIa inhibitor. A randomized, double-blind clinical trial found that 5.4% of subjects in the angioplasty-with-eptifibatide group and 9.2% in the angioplasty-with-

placebo group had either died or experienced another heart attack following treatment.[15] *Exercises 20.19 and 20.20 are based on this study.*

20.19 (Optional) What is the relative risk reduction RRR$RRR$ in this study?

a. 0.587
b. 0.413
c. 0.704

20.20 (Optional) What is, approximately, the number needed to treat NNT NNT in this study?

a. 4
b. 9
c. 26

CHAPTER 20 EXERCISES

We recommend using the plus four method for all confidence intervals for proportions. However, the large-sample method is acceptable when the guidelines for its use are met.

20.21 Genetically altered mice. Genetic influences on cancer can be studied by manipulating the genetic makeup of mice. One of the processes that turn genes on or off (so to speak) in particular locations is called "DNA methylation." Do low levels of this process help cause tumors? Compare mice altered to have low levels of DNA methylation with normal mice. Of 33 mice with lowered levels of DNA methylation, 23 developed tumors. None of the control group of 18 normal mice developed tumors in the same time period.[16]

a. Explain why we cannot safely use either the large-sample confidence interval or the test for comparing the proportions of normal and altered mice that develop tumors.
b. The plus four method adds two observations, a success and a failure, to each sample. What are the sample sizes and the numbers of mice with tumors after you do this? Give a plus four 99% confidence interval for the difference in the proportions of the two populations that develop tumors.
c. Based on your confidence interval, is the difference between normal and altered mice significant at the 1% level?

20.22 Shrubs that withstand fire. Some shrubs can resprout from their roots after their tops are destroyed. Fire is a serious threat to shrubs in dry climates, as it can injure the roots as well as destroy the tops. One study of resprouting took place in a dry area of Mexico.[17] The investigators randomly assigned 12 shrubs to the treatment group and 12 shrubs to the control group. They clipped the tops of all the shrubs. They then applied a propane torch to the stumps of shrubs in the treatment group to simulate a fire. All the shrubs in the treatment group eventually resprouted but only 8 in the control group did.

a. We want to estimate the difference in success rates for the two approaches. We should not use the large-sample confidence interval, however. Why not?

b. The plus four method adds two observations, a success and a failure, to each sample. What are the sample sizes and the numbers of resprouting shrubs afterward?

c. Obtain the plus four 95% confidence interval for the difference between the proportion of resprouting shrubs in dry areas when treated with fire or not. Interpret your interval in context.

20.23 Treating AIDS. The drug AZT was the first drug that seemed effective in delaying the onset of AIDS in HIV-positive patients. Evidence for AZT’s effectiveness came from a large randomized comparative experiment. The subjects were 1300 HIV-positive volunteers who had not yet developed AIDS. The study assigned 435 of the subjects at random to take 500 milligrams of AZT each day, and another 435 subjects to take a placebo. (The others were assigned to a higher dose of AZT, but we will compare only the first two groups.) At the end of the study, 38 of the placebo recipients and 17 of the AZT recipients had developed AIDS. We want to test the claim that taking AZT lowers the proportion of infected people who will develop AIDS in a given period of time.

a. State the hypotheses and check that you can safely use the z procedures.

b. How significant is the evidence that AZT is effective?

c. The experiment was double-blind. Explain what this means.

(*Comment:* Medical experiments on treatments for AIDS and other fatal diseases raise hard ethical questions. Some people argue that because AIDS is always fatal, infected people should get any drug that has any hope of helping them. The counter-argument is that then we will never find out which drugs really work. The placebo recipients in this study were given AZT as soon as the results indicated that AZT was clearly more effective.)

20.24 Peanut allergies in children. Peanut allergy is increasingly common in Western countries. A study enrolled infants with a diagnosed peanut sensitivity and randomly assigned them to either avoid or regularly consume peanuts until 60 months of age. At the end of the study, 18 of the 51 children who had avoided peanuts were still allergic to peanuts. In contrast, 5 of the 47 children who had regularly consumed peanuts were still allergic to peanuts.[18] We want to run a test to see if peanut allergy at 60 months is affected by the long-term consumption or avoidance of peanuts among children diagnosed in infancy with a peanut allergy.

a. Check the conditions for inference and state the hypotheses.
b. Obtain the test statistic and *P*-value.
c. Conclude in context.

20.25 Arthroscopic surgery. Arthroscopic surgery for the treatment of knee pain is one of the most common orthopedic procedures, despite a lack of relevant scientific studies. One double-blind study randomly assigned patients with knee pain to arthroscopic surgery or sham surgery. One year later, 62 of the 70 patients in the treatment group reported improvement, compared with 63 of the 76 patients in the control group.[19] Do the data give good evidence in support of arthroscopic surgery for the treatment of knee pain?

a. Check the conditions for inference and state the hypotheses.
b. Obtain the test statistic and *P*-value.
c. Conclude in context.

20.26 Peanut allergies in children, continued. Go back to Exercise 20.24. We want to estimate how much difference the two approaches make by using a confidence interval for the difference between the proportions with a peanut allergy at 60 months in the two populations.

a. Explain why we cannot use the large sample method.
b. Obtain the appropriate 95% plus four confidence interval.
c. Write a couple of sentences describing the study findings in context. Be sure to cite your confidence interval.

20.27 Staph infections. Exercise 19.41 (page 492) described a study of staph infections in surgery patients. The researchers recruited 917 patients who had tested positive for *Staphylococcus aureus* and randomly assigned them to receive a nasal ointment that contained either a staph-killing solution or a placebo. Postsurgery infections with *S. aureus* were then recorded. In the treatment group, 17 of 504 patients developed a staph infection, compared with 32 of 413 patients in the placebo group.[20]

a. Is there good evidence that treatment of staph-positive patients before surgery reduces the rate of postsurgical staph infection? State the null and alternative hypotheses, and obtain the test statistic and *P*-value.
b. Calculate a 95% confidence interval for the difference between the proportions of postsurgical staph infections among patients receiving a staph-killing ointment or a placebo.

20.28 Online harassment. People who experience online harassment either ignore it or respond via postings or involving authorities. Is one method more effective than the other? A 2014 Pew Research Center report, based on a random sample of American adults, found that 412 of the 549 respondents who had chosen to ignore online harassment thought that it was an effective way to deal with the issue. Of the 368 respondents who had chosen to respond to online harassment, 305 thought that it had been effective.

a. Is there good evidence that one approach is more likely to be effective than the other? State the null and alternative hypotheses, obtain the test statistic and *P*-value, and conclude in context.
b. How large is the difference? Obtain and interpret in context a 95% confidence interval for the difference between the proportions of American adults who find ignoring online harassment effective and those who find responding to it effective.
c. How effective is that best method? Obtain and interpret in context a 95% confidence interval for the proportion of American adults who find the best method effective.

20.29 Children’s food choices. Exercise 19.39 (page 492) describes a study in which 40 young children were asked to taste Graham Crackers presented in a package either with or without a popular cartoon character. When asked to indicate which of the two options they would prefer to eat for a snack, 35 chose the version with a cartoon on the package. When asked to choose between packages of baby carrots with or without a popular cartoon character, 29 said they would prefer to eat the version with a cartoon on the package. Explain why we *cannot* use the methods introduced in this chapter to compare the proportions of young children selecting the packaging with a popular cartoon when the food item is either Graham Crackers or baby carrots.

In responding to Exercises 20.30 to 20.40, follow the **Plan, Solve,** *and* **Conclude** *steps of the four-step process. It may be helpful to restate in your own words the* **State** *information given in the exercise.*

20.30 Knuckle cracking. Urban legend says that knuckle cracking is bad for you and will eventually lead to osteoarthritis of the hand joints. A retrospective case-control study of individuals who had received an X-ray of the right hand in the past five years enrolled 135 subjects with a confirmed diagnosis of hand osteoarthritis and 80 control subjects in the same age group. The study found that 24 subjects with hand osteoarthritis and 19 control subjects were habitual knuckle crackers.[21] Is there good evidence that the proportion of knuckle crackers is greater among individuals with hand osteoarthritis? Does this support the urban legend about knuckle cracking?

20.31 Altruism in prairie dogs. Is altruistic behavior influenced by its potential cost? Prairie dogs are social rodents who warn one another of a predator's presence with barking calls. A researcher examined whether close proximity to a predator results in less frequent alarm calls because the callers bear more personal risk. The "predator"—a stuffed badger controlled remotely—was placed multiple times randomly either near or far from a prairie dog colony. The number of warning calls made by isolated foraging colony members is displayed in the following table:[22]

	Distance to Predator	
	Near	**Far**
Alert call	29	55
No call	80	78

Do the data support the researcher's hypothesis?

20.32 More on altruism in prairie dogs. The researcher in the Exercise 20.31 also hypothesized that if prairie dogs make fewer calls when they feel individually more at risk, it will be reflected in their behavior. That is, individuals remaining aboveground will make a higher proportion of alert calls than individuals seeking cover upon appearance of a predator. Here are the study's findings:

	Behavioral Response	
	Seek cover	**Stay above ground**
Alert call	11	95
No call	63	121

Do the data support the researcher's second hypothesis?

20.33 Did the random assignment work? A large clinical trial of the effect of diet on breast cancer assigned women at random to either a normal diet or a low-fat diet. To check that the random assignment did produce comparable groups, we can compare the two groups at the start of the study. When asked if they had a family history of breast cancer, 3396 of the 19,541 women in the low-fat group and 4929 of the 29,294 women in the control group said "yes."[23] If the random assignment worked well, there should *not* be a significant difference in the proportions with a family history of breast cancer. How significant is the observed difference?

20.34 Aflatoxicosis in Kenya. Aflatoxins are toxic compounds secreted by fungi that are found most often in damaged crops. Kenya experienced an outbreak of aflatoxicosis in 2004 that resulted in several hundred cases of liver failure in humans, including 125 deaths. The Kenyan Ministry of Health suspected that improper maize (corn) storage conditions were at least partly responsible for the outbreak. A random sample of 27 case-patients with aflatoxicosis and 43 healthy controls were asked whether they stored their maize in the house or in a dedicated granary. Prolonged storage in the house exposes crops to levels of humidity that can foster fungal growth. The study found that 22 of the case-patients and 23 of the controls had stored maize in the house.[24]

Choose carefully a valid statistical method to compare the effectiveness of both storage methods against aflatoxicosis. Do your results support the ministry’s hypothesis?

20.35 Surgical treatment of migraines. Migraine headaches are a neurological condition with a number of symptoms, including severe pain in half of the head that can last from 4 to 72 hours. A double-blind, sham-surgery–controlled study examined the effectiveness of facial surgery that removed small portions of migraine-triggering muscle or nerve tissue in an attempt to provide permanent relief from migraines. Seventy-five patients with moderate to severe migraines were randomly assigned to receive either actual or sham surgery in their predominant trigger site.[25]

a. A year later, 15 of the 26 subjects in the sham surgery group and 41 of the 49 subjects in the actual surgery group experienced at least a 50% reduction in migraine headaches. How significant is the observed difference between the proportions of subjects with an improved condition?
b. The study also reported that 28 of 49 patients in the actual surgery group reported complete elimination of migraines a year after treatment, compared with only 1 of 26 patients in the sham surgery group. How much better is the surgical treatment than the sham procedure in completely eliminating migraines? Give a 95% confidence interval using the plus four method.

20.36 Smoking cessation. Chantix is a medication prescribed for smoking cessation that works by targeting nicotine receptors in the brain. Chantix’s website describes a study of the drug efficacy.[26]

a. In its first stage, the study enrolled 1927 smokers who wanted to quit smoking and provided them with Chantix for twelve weeks. At the end of this first period, 1236 subjects had successfully quit smoking. Give a 95% confidence interval for the proportion of smokers using Chantix for twelve weeks who successfully quit smoking by the end of the twelve-week period.
b. In the second stage of the study, the 1236 subjects who had

successfully quit smoking after twelve weeks of Chantix were randomly assigned to remain on Chantix or to receive a placebo for another twelve weeks. At the end of this period, 425 of the 602 subjects given Chantix and 301 of the 604 subjects given a placebo remained tobacco free. Is there a significant difference in the proportions of ex-smokers who remain tobacco free in the two groups?

20.37 Duct tape for wart removal. A study compared the effectiveness of duct tape with cryotherapy with liquid nitrogen in the treatment of common warts in children and young adults. A total of 61 patients aged 3 to 22 years were randomly assigned to either treatment (duct tape applied directly on warts continuously for up to two months or up to six applications of liquid nitrogen every two or three weeks). Of the 26 patients treated with duct tape, 22 showed complete wart remission. In comparison, 15 of the 25 patients treated with cryotherapy achieved complete wart remission.[27]

a. Explain why the study cannot be double-blind.

b. Choose carefully a valid statistical method to compare the effectiveness of both methods. What do you conclude?

20.38 Do free samples influence prescription decisions? To answer this question, a study randomly assigned resident physicians to either accept free drug samples or refuse them. The researchers then examined prescriptions for drugs that are available both by prescription and in an over-the-counter (OTC) version. Of the 202 relevant prescriptions written by physicians accepting free samples, 51 stipulated the OTC version. Of the 188 relevant prescriptions written by physicians refusing free samples, 73 stipulated the OTC version.[28] Consider these data as two random samples from the populations of prescriptions written by resident physicians with or without access to free samples. Is there good evidence that physicians are influenced by free samples and prescribe proportionally fewer OTC versions when they receive free samples?

20.39 Carcinogenicity of electromagnetic fields. The U.S. National Toxicology Program studies the toxicity and carcinogenicity of agents potentially causing a risk to human health. Electromagnetic fields around electrical installations are harmless in theory, but they might have a long-term effect on health—for instance, in triggering tumors. Observational studies in human populations have suggested a potential association. One study looked at the occurrence of cancers in otherwise healthy rats after two years of daily exposure to 60-hertz electromagnetic fields. Several groups were studied, each exposed to a different electromagnetic intensity, as well as a control group kept in the same conditions but not exposed to any electromagnetic field. Here are the number of rats with a tumor after the two-year study period for the control group and for the group exposed to 2 gauss, which is approximately 1000-fold the intensity level considered high exposure for humans:[29]

Treatment	Male rats	Female rats
Control: no exposure	16 (n=99)$16\ (n = 99)$	19 (n=100)$19\ (n = 100)$
2-gauss exposure	30 (n=100)$30\ (n = 100)$	22 (n=100)$22\ (n = 100)$

In rats not exposed to an electromagnetic field, is there significant evidence of a difference in the tumor rates between males and females?

20.40 Carcinogenicity study, continued. Use the data from Exercise 20.39 to answer the following questions.

a. Is there significant evidence of a difference in the tumor rates in male rats either exposed to 2 gauss or not exposed?
b. Is there significant evidence of a difference in the tumor rates in female rats either exposed to 2 gauss or not exposed?
c. Write a brief summary of your findings from this exercise and Exercise 20.39.

20.41 Melatonin and menopause. Melatonin is a naturally occurring hormone involved in the regulation of the body’s internal clock and is available as a dietary supplement in the United States. An experiment was designed to assess the effect of nighttime melatonin intake on the levels of luteinizing hormone (LH) in perimenopausal and menopausal women. Women were randomly assigned to take either melatonin or a placebo daily for six months. Hormone levels were determined from blood samples taken at the beginning and at the end of the study. The response variable was whether or not a woman had an LH increase of at least 10%. Here are the results, by age group:[30]

Treatment	Women in their 40s	Women in their 50s and 60s
Melatonin	10 (n=38)$10\ (n = 38)$	22 (n=30)$22\ (n = 30)$
Placebo	20 (n=34)$20\ (n = 34)$	18 (n=38)$18\ (n = 38)$

a. For women in their 40s, is there significant evidence of a difference in the proportions of women with increased LH levels between women taking melatonin and women taking a placebo?

b. For women in their 50s and 60s, is there significant evidence of a difference in the proportions of women with increased LH levels between women taking melatonin and women taking a placebo?

c. Write a short description of your findings, contrasting your results for parts a and b.

d. If we had ignored age, the results would be 32 out of 68 women with increased LH in the melatonin condition compared with 38 out of 72 in the placebo condition. Show that these results are not statistically significant. This is an example of Simpson’s paradox, which was described in Chapter 5 (page 131).

The following exercises concern the optional material on relative risk and odds ratio in the health sciences.

20.42 Cardiovascular disease prevention. The Heart Outcomes Prevention Evaluation trial studied the effectiveness of angiotensin-converting enzyme (ACE) inhibitors in cardiovascular disease prevention. Part of the study randomly assigned diabetic patients older than age 55 to receive either the ACE inhibitor ramipril or a placebo daily for five years. Of the 1808 subjects taking ramipril daily, 112 died of a cardiovascular accident during the five-year study period. In contrast, 172 of the 1769 subjects who received a placebo died of a cardiovascular accident.[31]

a. Calculate the relative risk reduction and write a short statement using this measure to summarize the study findings.

b. Calculate the number needed to treat to save 1 life from cardiovascular death, and write a short statement using this measure to summarize the study findings.

c. Some studies also report the number of pills that must be taken to save 1 life. Calculate this value considering, for simplicity, that each subject in the treatment group took 1 pill every day for five years.

20.43 Treating AIDS. Go back to Exercise 20.23, which describes the first study of AZT effectiveness in delaying the onset of AIDS.

a. Calculate and interpret the odds ratio and the relative risk for the AZT group compared with the placebo group. How do they compare?

b. Calculate and interpret the absolute and the relative risk reduction for AIDS onset.

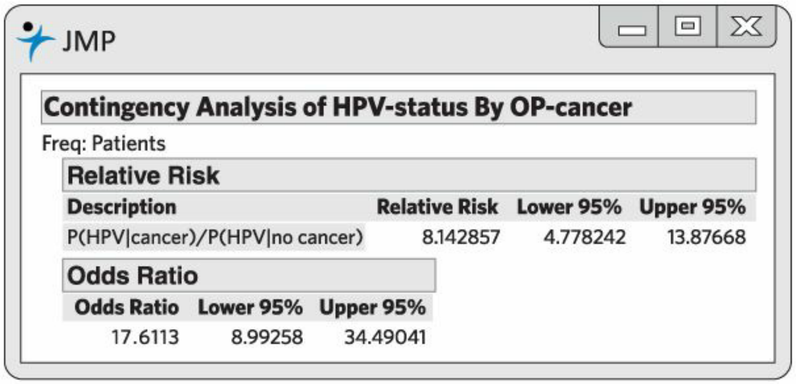
c. Calculate the number needed to treat with AZT compared with a placebo to prevent the onset of AIDS in 1 patient.

20.44 HPV and oropharyngeal cancer. Molecular evidence suggests a link between the human papillomavirus (HPV) and oropharyngeal cancer. An epidemiologic study enrolled 100 hospitalized patients with newly diagnosed oropharyngeal cancer and 200 control hospitalized patients without oropharyngeal cancer. Oral HPV infection was strongly associated with oropharyngeal cancer, with 57 of the 100 cancer patients having an HPV infection, compared with only 14 of the 200 patients without cancer.[32] Figure 20.5 shows the JMP output giving 95% confidence intervals for the relative risk and odds ratio for these data, similar to the one provided for Example 20.6.

a. What are the proportions of patients with an HPV infection in the two groups?

b. What is the relative risk of having an HPV infection for hospitalized patients with or without newly diagnosed oropharyngeal cancer? Interpret the corresponding confidence interval in context.

c. What is the odds ratio for HPV infection, comparing patients with and patients without newly diagnosed oropharyngeal cancer? Interpret the corresponding confidence interval in context.



```
Contingency Analysis of HPV-status By OP-cancer
Freq: Patients
  Relative Risk
  Description                      Relative Risk   Lower 95%   Upper 95%
  P(HPV|cancer)/P(HPV|no cancer)        8.142857    4.778242    13.87668
  Odds Ratio
   Odds Ratio  Lower 95%  Upper 95%
      17.6113    8.99258   34.49041
```

Figure 20.5

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 20.5 JMP output for a study of the association between HPV and oropharyngeal cancer.

CHAPTER 21 The Chi-Square Test for Goodness of Fit

Simon Fraser/Science Source

IN THIS CHAPTER WE COVER...

- Hypotheses for goodness of fit
- Expected counts and chi-square statistic
- The chi-square test for goodness of fit
- Interpreting significant chi-square results
- Conditions for the chi-square test

In Chapter 19 we used the one-sample z procedures to obtain confidence intervals and test hypotheses about the proportion of successes in a population. The definition of “success” was arbitrary and simply referred to the outcome of interest, with any other outcome constituting a “failure.” Such dichotomous labeling is sometimes an obvious choice, as in Example 19.6 (page 486), in which we examined whether or not live aphids land on their legs after a drop. In other situations, a more detailed description of possible outcomes would be helpful. For instance, we might want to study the response to a new treatment by counting the number of subjects whose health status improved, remained stable, or deteriorated. In this chapter we introduce a new statistical method, the chi-square test, that allows us to test hypotheses about one categorical variable with two *or more* levels. We will see in Chapter 22 that this versatile method can also be adapted to study the relationship between two independent categorical variables.

HYPOTHESES FOR GOODNESS OF FIT

When a species has several variants of a phenotype passed on from generation to generation, we can form a hypothesis about the genetics of the trait based on Mendelian theories of genetic inheritance. For example, in a two-gene dominant epistatic model, the first gene masks the effect of the second gene, leading to the expression of three phenotype variants. Crossing the dominant and recessive homozygote lines would result in a second generation represented by a mix of dominant, intermediate, and recessive phenotype variants in the following expected proportions: 12/16, 3/16, and 1/16, respectively (also called a 12:3:1 ratio). Such a model can provide the basis for the null hypothesis in a significance test.

EXAMPLE 21.1 Genetics of seed color

Cultivated amaranth grains (*Amaranthus caudatus*) come in three colors: black, brown, and pale. Geneticists asked whether this phenotype could result from a dominant epistatic control. Crossing black-seeded and pale-seeded *A. caudatus* populations (the homozygote lines) gave the following counts of black, brown, and pale seeds in the second generation (F2):[1]

Seed coat color	black	brown	pale
Seed count	321	77	31

We want to know whether seed color in *A. caudatus* could arise from a dominant epistatic genetic control. Mendelian genetics provides a corresponding model of expected proportions in the F2 population that we use to set up the null hypothesis:

H0: pblack=1216=34 and pbrown=316 and ppale=116

$$H_0 : p_{black} = \frac{12}{16} = \frac{3}{4} \quad \text{and} \quad p_{brown} = \frac{3}{16} \quad \text{and} \quad p_{pale} = \frac{1}{16}$$

The alternative hypothesis is simply that H0H_0 is not true.

The null hypothesis describes the hypothesized distribution of a categorical variable in the target population. The statistical test we will use for that purpose is the chi-square test for *goodness of fit*. The underlying idea is that the test assesses whether the observed counts "fit" the distribution outlined by H_0. The alternative hypothesis for a goodness-of-fit test is always simply that the null hypothesis is not true.

In Example 21.1 the variable is seed color and the population of interest is all seeds in the F2 generation of a homozygotic dominant–recessive cross; the hypothesized distribution is derived from genetic theory. *Avoid the temptation of stating the null hypothesis in terms of the sample proportions. This is a mistake often made by students. The null hypothesis must reflect your assumptions, not the data used to test it.* Here is an example with a different approach in which we are asking whether the outcomes may simply be the result of chance.

CAUTION

EXAMPLE 21.2 Righting behavior of aphids

Pea aphids (*Acyrthosiphon pisum*) are small, wingless, sap-sucking insects that live on plants. They evade predators (such as ladybugs) by dropping off. A study examined the mechanism of aphid drops. Researchers hung live aphids upside down from delicate tweezers and then released them. The videotaped drops showed that 19 of the 20 aphids landed on their legs.[2] Is this evidence that live aphids land right side up (on their legs) more often than chance alone would predict?

In this scenario, aphids can land either on their legs or on their back. If chance alone was at work, we would expect aphids to be equally likely to land on either side. Therefore, the relevant null hypothesis is

$$H_0: p_{legs} = \frac{1}{2} \quad \text{and} \quad p_{back} = \frac{1}{2}$$

The alternative hypothesis is simply that H_0 is not true.

Simon Fraser/Science Source

The goodness-of-fit test can assess a variety of null hypotheses that reflect how we expect one categorical variable to be distributed in the target population. Because the variable is categorical, the parameters in H0 H_0 are population proportions for the kk possible outcomes (levels) making up the variable. The only constraint on the null hypothesis for a goodness-of-fit test is that these kk proportions sum to 1, so that H0H_0 accounts for all possible outcomes. In Example 21.1 the variable seed color had three possible outcomes (black, brown, pale) and the three population proportions in H0H_0 were 12/16, 3/16, and 1/16, which sum to 1. In Example 21.2 the variable landing side had only two possible outcomes (legs, back) and the two population proportions in H0H_0 were 1/2 and 1/2.

APPLY YOUR KNOWLEDGE

21.1 Saving birds from windows. Many birds are injured or killed by flying into windows. It appears that birds don’t see windows. Can tilting windows down so that they reflect earth rather than sky reduce bird strikes? Researchers placed six windows at the edge of a woods: two vertical, two tilted downward 20 degrees, and two tilted downward 40 degrees. During the next four months, there were 53 bird strikes: 31 on the vertical windows, 14 on the 20-degree windows, and 8 on the 40-degree windows.[3] Does the tilt have an effect? State the null and alternative hypotheses.

a. Identify the variable and explain how you know that it is categorical. How many levels does this variable have?

b. State the null and alternative hypotheses that are appropriate for the study’s objective.

Carol Bloomfield/Alamy Stock Photo

21.2 Heat resistance in rice. Rice is one of the most widely consumed grains worldwide. As increasing global temperatures hamper rice production, scientists are searching for varieties exhibiting some degree of heat resistance. The HT54 variety of the indica rice, *Oryza sativa*, was found to tolerate several days of high temperatures up to 48 degrees Celsius (118 degrees Fahrenheit) during its growth. Researchers performed a genetic crossing experiment between the heat-resistant HT54 variety and a control variety, HT13, without heat resistance. All seedlings in the first generation survived a heat challenge of 48 degrees Celsius. In the second generation (F2), only 548 of the 744 plants survived the heat challenge.[4] This would suggest that heat resistance is carried by a single gene with a dominant allele and a recessive allele. Are the findings consistent with a simple dominant–recessive Mendelian genetic model in which F2 is made up of 75% dominant-trait (heat resistance) and 25% recessive-trait rice plants?

a. Identify the variable and explain how you know that it is categorical. How many levels does this variable have?

b. State the null and alternative hypotheses that correspond to the study’s objective.

EXPECTED COUNTS AND CHI-SQUARE STATISTIC

We have stated the null and alternative hypotheses, and now we need to test whether the observed results differ significantly from our expectations under H_0. To test H_0, we compare the observed counts with the *expected counts*, the counts we would expect—except for random variation—if H_0 was true. When the observed counts are far from the expected counts, that is evidence against H_0.

The expected count, in n independent trials, for a particular outcome with probability p of occurring is simply np. (Chapter 12 explains why in detail.) This doesn't mean that we expect to see exactly np counts of that outcome in n trials. Rather, we expect to see np counts on average over the long run if we repeat the n trials many, many times.

EXPECTED COUNTS

A categorical variable has k possible outcomes, with probabilities $p_1, p_2, p_3, \ldots, p_k$ of occurring. That is, p_i is the probability of the ith outcome. We have n independent observations from this categorical variable.

To test the null hypothesis that the probabilities have specified values

H0: p1=p10, p2=p20, …, pk=pk0

$$H_0: p_1 = p_{1_0},\ p_2 = p_{2_0},\ \ldots,\ p_k = p_{k_0}$$

first compute an **expected count** for each of the k outcomes as follows:

expected count of outcome i=npi0

$$\text{expected count of outcome } i = np_{i_0}$$

EXAMPLE 21.3 **Genetics of seed color: expected counts**

Example 21.1 described the color of 429 second-generation (F2) *A. caudatus* seeds after crossing black-seeded and pale-seeded populations. The null hypothesis, based on a dominant epistatic model of genetic inheritance, assumed that

H0: pblack=1216=34 and pbrown=316 and ppale=116

$$H_0: p_{black} = \frac{12}{16} = \frac{3}{4} \quad \text{and} \quad p_{brown} = \frac{3}{16} \quad \text{and} \quad p_{pale} = \frac{1}{16}$$

If H_0 was true, we would expect the distribution of seed color to be

black seeds: npblack0=429×3/4=321.75brown seeds: npbrown0=429×3/16=80.4375pale seeds: nppale0=429×1/16=26.8125

$$\begin{aligned} \text{black seeds}: np_{black\,0} &= 429 \times 3/4 = 321.75 \\ \text{brown seeds}: np_{brown\,0} &= 429 \times 3/16 = 80.4375 \\ \text{pale seeds}: np_{pale\,0} &= 429 \times 1/16 = 26.8125 \end{aligned}$$

alex5248/Thinkstock

Expected counts are hypothetical counts and do not have to be round numbers. However, because the population proportions p_i under H_0 must sum to 1, the expected counts must sum to the total number of observations n in the random sample. That is, the expected counts are simply a statement of how these n observations would be expected to fall into the k possible outcomes (levels) of the categorical variable, according to H_0.

CAUTION

To compare observed and expected counts in a goodness-of-fit test and ask whether they differ significantly, we need a test statistic. That test statistic is the *chi-square statistic,* X^2.

CHI-SQUARE STATISTIC

The **chi-square statistic** is a measure of how far the observed counts in a random sample are from the expected counts defined by the null hypothesis. The formula for the statistic is

X2=∑(observed count - expected count)2expected count=(observed1-expected1)2expected1+···+(observedk-expectedk)2expectedk

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = \frac{(\text{observed}_1 - \text{expected}_1)^2}{\text{expected}_1} + \cdots + \frac{(\text{observed}_k - \text{expected}_k)^2}{\text{expected}_k}$$

where kk is the number of different outcomes that the categorical variable can take. Each of the kk terms in the sum is called a **chi-square component.**

Think of the chi-square statistic, X2X^2, as a measure of the relative distance of the observed counts from the expected counts. Like any distance, it is always either zero or positive, and it is zero only when the observed counts are exactly equal to the expected counts. Small values of X2X^2 represent small deviations from H0H_0 that do not provide sufficient evidence to reject H0H_0. Conversely, large values of X2X^2 are evidence against H0H_0, because they indicate that the observed counts are far from what we would expect if H0H_0 was true. *The alternative hypothesis,* HaH_a, *for the chi-square test is nonspecific, or nondirectional, because any violation of* H0H_0 *tends to produce a large value of* X2X^2.

CAUTION

EXAMPLE 21.4 Genetics of seed color: calculating X^2

In Example 21.3 we obtained the expected counts of *A. caudatus* seeds of various color under a dominant epistatic model of genetic inheritance.

Seed coat color	black	brown	pale
Observed count	321	77	31
Expected count	321.75	80.4375	26.8125

We can now calculate the chi-square statistic to test H0H_0.

X2=∑(observed count - expected count)2expected count=(321-321.75)2321.75+(77-80.4375)280.4375+(31-26.8125)226.8125=0.8026

$$\begin{aligned} X^2 &= \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \\ &= \frac{(321-321.75)^2}{321.75} + \frac{(77-80.4375)^2}{80.4375} + \frac{(31-26.8125)^2}{26.8125} \\ &= 0.8026 \end{aligned}$$

APPLY YOUR KNOWLEDGE

21.3 Saving birds from windows, continued. Exercise 21.1 described an experiment designed to figure out whether tilting windows down so that they reflect earth rather than sky can help reduce accidental bird strikes. Calculate the expected counts in each of the three conditions and compute the chi-square statistic.

21.4 Heat resistance in rice, continued. Exercise 21.2 investigated a genetic model of heat resistance in rice plants. What are the expected counts for each of the two phenotypes? Use these expected counts to compute the chi-square statistic.

THE CHI-SQUARE TEST FOR GOODNESS OF FIT

Hypothesis tests require a set of null and alternative hypotheses, a test statistic, and a *P*-value that gives the probability if the null hypothesis was true of obtaining a test statistic at least as extreme as that computed. To obtain the *P*-value for a goodness-of-fit test, we need to have a known probability distribution for the test statistic, X2χ^2. Deriving this distribution mathematically is beyond the scope of an introductory statistics textbook, so we will simply state the facts. When the null hypothesis is true, the X2χ^2 statistic follows a *chi-square distribution.*

THE CHI-SQUARE DISTRIBUTIONS

The **chi-square distributions** are a family of distributions that take only positive values and are skewed to the right. A specific chi-square distribution is specified by giving its **degrees of freedom.**

Figure 21.1 shows the density curves for three members of the chi-square family of distributions. As the degrees of freedom increase, the density curves become less skewed and larger X2χ^2 values become more probable. Because chi-square distributions take only positive values, the more extreme values are always found on the right tail of the curve.

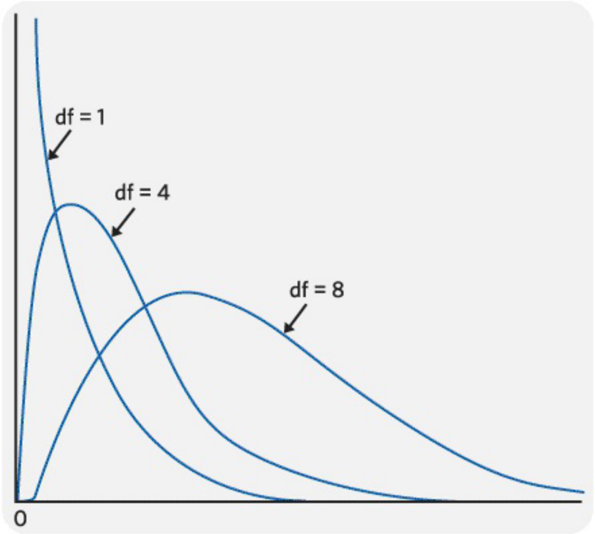


Figure 21.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 21.1 Density curves for the chi-square distributions with 1, 4, and 8 degrees of freedom. Chi-square distributions take only positive values and are right-skewed.

P-VALUE OF THE CHI-SQUARE GOODNESS-OF-FIT TEST

In a chi-square goodness-of-fit test involving kk **outcomes,** when the null hypothesis is true, the test statistic X2χ^2 has the chi-square distribution with k-1$k-1$ **degrees of freedom.**

The *P*-value of the chi-square test is the area to the right of the calculated X2χ^2 statistic under this chi-square distribution.

Software usually finds *P*-values for us. Table D in the back of the book gives critical values for several chi-square distributions and can be used to find approximate *P*-values when no software is available.

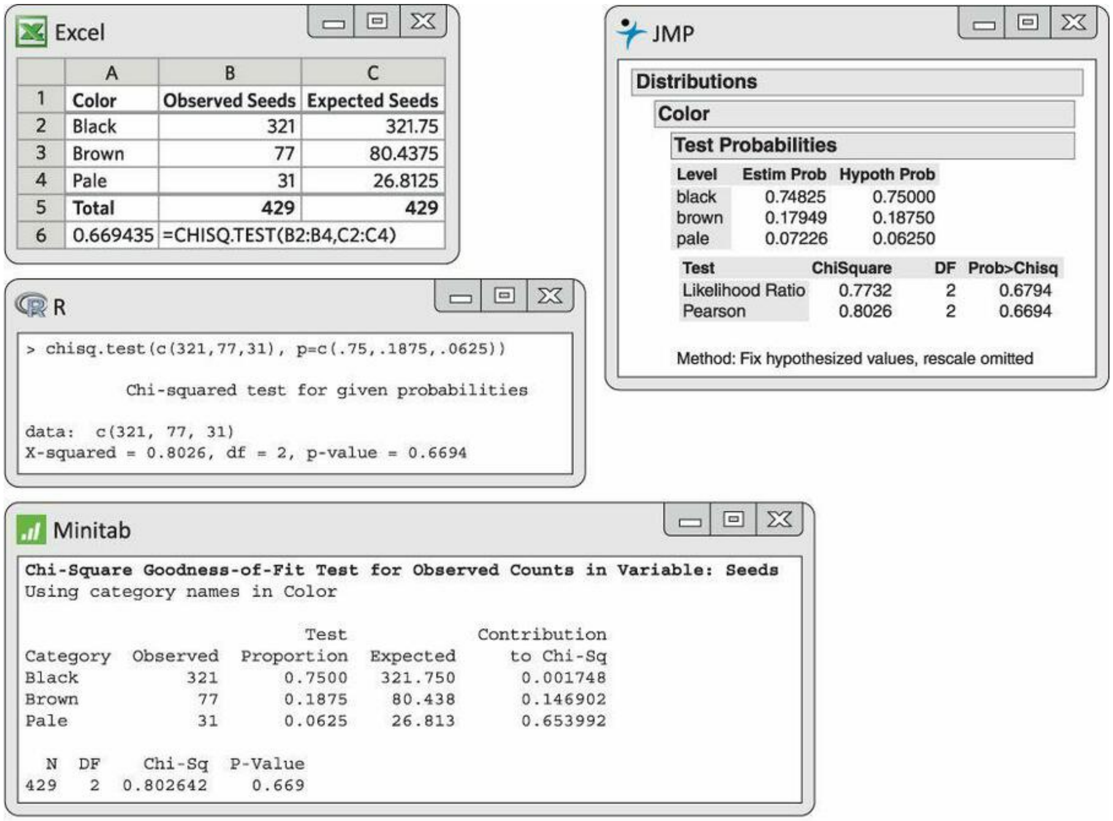
EXAMPLE 21.5 Genetics of seed color: *P*-value and conclusion

Example 21.1 described the seed color of 429 second-generation (F2) amaranth plants after crossing black-seeded and pale-seeded populations. The null hypothesis, based on a dominant epistatic model of genetic inheritance, assumed that

H0: pblack=1216=34 and pbrown=316 and ppale=116

$$H_0 : p_{black} = \frac{12}{16} = \frac{3}{4} \quad \text{and} \quad p_{brown} = \frac{3}{16} \quad \text{and} \quad p_{pale} = \frac{1}{16}$$

Figure 21.2 shows outputs for the corresponding chi-square goodness-of-fit test from three statistical programs and a spreadsheet program. The test statistic is X2=0.8026$X^2 = 0.8026$ and the test *P*-value is 0.669.This *P*-value is not statistically significant, so we cannot reject the null hypothesis.



Excel

	A	B	C
1	Color	Observed Seeds	Expected Seeds
2	Black	321	321.75
3	Brown	77	80.4375
4	Pale	31	26.8125
5	Total	429	429
6	0.669435	=CHISQ.TEST(B2:B4,C2:C4)	

R

```
> chisq.test(c(321,77,31), p=c(.75,.1875,.0625))

          Chi-squared test for given probabilities

data:  c(321, 77, 31)
X-squared = 0.8026, df = 2, p-value = 0.6694
```

JMP

Distributions

Color

Test Probabilities

Level	Estim Prob	Hypoth Prob
black	0.74825	0.75000
brown	0.17949	0.18750
pale	0.07226	0.06250

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	0.7732	2	0.6794
Pearson	0.8026	2	0.6694

Method: Fix hypothesized values, rescale omitted

Minitab

```
Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Seeds
Using category names in Color

                             Test                Contribution
Category  Observed     Proportion   Expected      to Chi-Sq
Black          321         0.7500    321.750       0.001748
Brown           77         0.1875     80.438       0.146902
Pale            31         0.0625     26.813       0.653992

  N  DF    Chi-Sq  P-Value
429   2  0.802642    0.669
```

Figure 21.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 21.2 Output from JMP, Minitab, R, and Excel for a study of amaranth seed color.

The findings are *not inconsistent* with the distribution of seed color predicted in F2 for a dominant epistatic genetic model (P=0.67)$(P = 0.67)$. Therefore, the genetic determination of amaranth seed color *could* indeed represent a case of dominant epistasis.

The outputs in Figure 21.2 differ in the type and amount of information they give. All but Excel provide the chi-square statistic, X2=0.8026 $X^2 = 0.8026$ (which JMP calls “Pearson chi-square”), computed as shown in

Example 21.4. Minitab additionally lists the χ^2 components in the column labeled “Contribution to Chi-sq.” All four outputs report a *P*-value rounding to 0.669.

The *P*-value 0.669 is the area to the right of $\chi^2 = 0.8026$ under the chi- square distribution with $k - 1 = 2$ degrees of freedom. The R command pchisq(0.8026, 2, lower.tail=FALSE) and the TI-83 calculator command X2cdf(0.8026,1E99,2) both confirm that this area is indeed equal to 0.669. If you do not have access to statistical software, you can use Table D to obtain an approximate test *P*-value. The value $\chi^2 = 0.8026$ is smaller than the smallest critical value, 2.77, reported on the $df = 2$ row of Table D, which indicates that the test *P*-value is greater than 0.25.

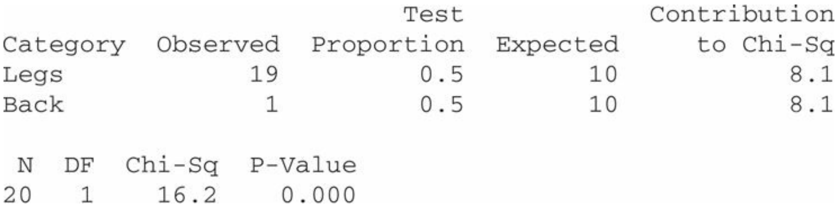
Despite its name, the “goodness-of-fit test” does not look for evidence that the data fit the null hypothesis. The fact that we fail to reject H_0 never implies that H_0 is true. We cannot claim that amaranth seed color is actually based on a dominant epistatic genetic model, only that such a model is plausible. That is, the data gathered are *consistent with* the 12:3:1 ratio in F2 that would result from dominant epistasis. Always remember that a large *P*-value can arise under any of the following conditions:

CAUTION

- H_0 is indeed true. Only further scientific investigation can confirm this.
- H_0 is not actually true but it is too close to the real population distribution for us to tell them apart statistically.
- H_0 is definitely not true but the sample size is too small or the variability too great to reach statistical significance.

Also keep in mind that H_0 only specifies the proportions of black, brown, and pale amaranth seeds in F2 of a genetic crossing. Even if H_0 was true, these proportions are not proof that the *cause* of this distribution of seed color would necessarily be dominant epistasis. Other genetic patterns, either alone or in combination with environmental factors, could potentially produce a 12:3:1 ratio in F2. The issue of causation in this study is not directly addressed by the chi-square test for goodness of fit.

EXAMPLE 21.6 Righting behavior of aphids, continued

> In Example 21.2 we described an experiment designed to examine the righting behavior of pea aphids, asking whether live aphids land right side up (on their legs) more often than chance alone would predict. Here is the Minitab output for a null hypothesis H0: plegs=12 and pback=12 $H_0 : p_{legs} = \frac{1}{2}$ and $p_{back} = \frac{1}{2}$, with observed counts of 19 aphids landing on their legs and 1 aphid landing on its back.
>
> 
>
> ```
>                                 Test                Contribution
> Category  Observed  Proportion  Expected    to Chi-Sq
> Legs            19         0.5        10          8.1
> Back             1         0.5        10          8.1
>
>  N  DF  Chi-Sq  P-Value
> 20   1    16.2    0.000
> ```
>
> The test *P*-value is so small that Minitab rounds it to 0.000 (indicating that P<0.0005$P < 0.0005$). This is highly statistically significant. The study provides very strong evidence that live aphids dropped from tweezers do not land on their legs or back at random.

The computations performed by Minitab in this example involve the following steps:

- Expected counts: nplegs0=20×0.5=10$np_{legs\,0} = 20 \times 0.5 = 10$ for legs, and npback0=20×0.5=10$np_{back\,0} = 20 \times 0.5 = 10$ for back.
- Test statistic:

  X2=∑(observed count - expected count)2expected count=(19-10)210+(1-10)210=8.1+8.1=16.2

  $$\begin{aligned} X^2 &= \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} \\ &= \frac{(19-10)^2}{10} + \frac{(1-10)^2}{10} \\ &= 8.1 + 8.1 = 16.2 \end{aligned}$$

- Degrees of freedom of the chi-square distribution: k-1=2-1=1 $k - 1 = 2 - 1 = 1$.
- *P*-value as the area to the right of X2=16.2$X^2 = 16.2$ under the chi-square distribution with df=1$df = 1$. Software commands such as X2cdf(16.2,1E99,1) in the TI-83 calculator or =CHISQ.DIST.RT(16.2,1) in Excel give a *P*-value of 5.7E−5 (or

0.000057). Table D tells us only that the *P*-value must be less than 0.0005 because the test statistic $\chi^2 = 162$. is greater than the largest critical value 12.12 listed in the $df = 1$ row.

If you studied Chapter 19, you should see that a one-sample z test for a proportion may also be used for the data in Example 21.6. This is because the categorical variable in that example (landing side) can take only one of two values—typically arbitrarily labeled success or failure. The null hypotheses tested are

$$H_0 : p_{success} = p_0$$

for theone-sample z test, and

$$H_0 : p_{success} = p_0 , p_{failure} = 1 - p_0$$

for the chi-square goodness-of-fit test.

These two tests always agree. In fact, the chi-square statistic χ^2 is just the square of the z statistic, and the *P*-value for χ^2 is exactly the same as the two-sided *P*-value for z. The main advantages of the one-sample z test are that you can select a one-sided alternative and the test it is related to a confidence interval for the population proportion of success p. The main advantage of the chi-square test for goodness of fit is its flexibility, such as its ability to work for categorical variables with more than just two possible outcomes.

The chi-square test for goodness of fit can also be used to test whether some data might follow a known probability distribution, such as the binomial or Poisson distributions described in Chapter 12. Using the chi-square test for this purpose first requires estimating the distribution parameters from the sample data. For each parameter that must be estimated this way (such as the probability of success p for a binomial distribution, or the mean number of occurrences μ for a Poisson distribution), the chi-square test for goodness of fit loses an additional degree of freedom. Exercise 21.36 walks you through the process step by step.

STATS IN YOUR WORLD

BIRD FATALITIES

According to the U.S. Fish and Wildlife Services (FWS), more birds die because of window collisions than because of any other factor associated with human activity. The FWS estimates that anywhere from 100 million to 1 billion birds are killed by window strikes each year, representing up to 5% of the North American fall bird population. Actual deaths are likely underestimated by the general public because some injured birds fly away before dying from injuries sustained and natural predators feed on injured and dying birds.

APPLY YOUR KNOWLEDGE

21.5 Saving birds from windows, continued. Exercise 21.1 described an experiment designed to figure out whether tilting windows down so that they reflect earth rather than sky could help reduce accidental bird strikes. Use software or Table D to obtain the test *P*-value, then conclude in context.

21.6 Heat resistance in rice, continued. Exercise 21.2 investigated a genetic model of heat resistance in rice plants. Software gives a *P*-value of 0.397. Interpret this result. Can you conclude that the distribution of heat-resistant phenotypes in F2 follows a dominant–recessive model (75%–25%)?

INTERPRETING SIGNIFICANT CHI-SQUARE RESULTS

The chi-square test is the overall test for detecting departures from a distribution model assumed under H0H_0. **When the test is statistically significant,** it is important to look at the data to understand the nature of the distribution. Here are three ways to look at the data after statistical significance has been established:

- **Compare appropriate percents:** Which outcomes occur in percents quite different from those hypothesized in H0?H_0?
- **Compare observed and expected counts:** Which outcomes have more or fewer observations than we would expect if H0H_0 was true?
- **Look at the chi-square components:** Which outcomes contribute the most to the value of X2?X^2?

The study of aphid righting behavior described in Example 21.6 was highly statistically significant. The researchers found that 19 aphids landed on their legs when the expected count under the null hypothesis was 10; inversely, only 1 aphid landed on its back when we would have expected from chance alone 10 aphids to do so. This tells us that live aphids dropped from tweezers are *more likely than chance* to land on their legs, suggesting a “righting” mechanism is at work.

STATS IN YOUR WORLD

WEEKEND OUTING?

You work all week—and then it rains on the weekend. Can there really be a statistical truth behind our perception that the weather is against us? At least on the East Coast of the United States, the answer is “yes.” Going back to 1946, it seems that Sundays receive 22% more precipitation than Mondays. The likely explanation is that the pollution from all those workday cars and trucks forms the seeds for raindrops—with just enough delay to cause rain on the weekend.

EXAMPLE 21.7 Not on the weekend, please

A random sample of 700 births from local records shows this distribution across the days of the week:

Day	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Births	110	124	104	94	112	72	84

STATE: Nature shouldn’t favor any particular day of the week, but modern births may not strictly follow the laws of nature. In fact, we can see that the two smallest counts of births in this sample are for Saturday and Sunday. Do these data give significant evidence that local births are not equally likely on all days of the week?

PLAN: Carry a chi-square test for goodness of fit on the distribution of the variable day of birth. The null hypothesis says that births *are* evenly distributed over the 7 days of the week. In that case, each day would get 1/7th of all births. So the null hypothesis is

H0: p1=p2=p3=p4=p5=p6=p7=17

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$$

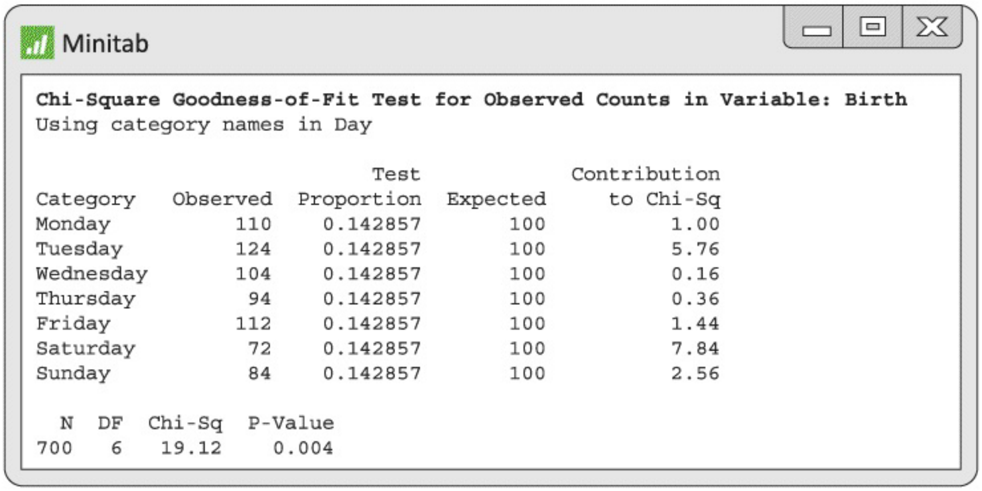
representing the discrete probability distribution for day of birth such that:

Day	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Probability	p1p_1	p2p_2	p3p_3	p4p_4	p5p_5	p6p_6	p7p_7

The alternative hypothesis simply says that days are *not* all equally likely for births (H0H_0 is not true, not all pi=17$p_i = \frac{1}{7}$):

SOLVE: **Check the conditions for inference.** We will need to get back to this part later in the chapter.

SOLVE: **Obtain the test statistic and *P*-value.** The Minitab output in Figure 21.3 shows that the chi-square statistic is X2=19.12$X^2 = 19.12$ and the *P*-value is 0.004.



```
Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Birth
Using category names in Day

                             Test              Contribution
Category    Observed   Proportion  Expected      to Chi-Sq
Monday           110     0.142857       100           1.00
Tuesday          124     0.142857       100           5.76
Wednesday        104     0.142857       100           0.16
Thursday          94     0.142857       100           0.36
Friday           112     0.142857       100           1.44
Saturday          72     0.142857       100           7.84
Sunday            84     0.142857       100           2.56

  N  DF  Chi-Sq  P-Value
700   6   19.12    0.004
```

Figure 21.3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 21.3 Minitab output for a study of birth days.

CONCLUDE: There is very strong evidence that the population distribution of local births is *not* evenly (“uniformly”) distributed across all 7 days of the week. The sample data are not consistent with a null hypothesis of equal proportions.

Because the test is statistically significant, we can look deeper to explain this finding. Look at the seven X2χ^2 components provided by Minitab in the rightmost column of Figure 21.3. Roughly 40% of the value of X2χ^2 (7.84 out of 19.12) comes from just one outcome. This points to the most important difference between the observations and the uniform model: a really low count of births on Saturday, representing the largest difference between any observed and expected counts. Most of the rest of X2χ^2 comes from two other outcomes: a higher than expected observed count of births on Tuesday and a lower than expected observed count of births on Sunday.

A similar conclusion can be reached after computing the percent of births in the sample that occurred on each day of the week.

Day	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Percent	16%	18%	15%	13%	16%	10%	12%

With 18% of the 700 births, Tuesday has nearly twice as many births as Saturday (only 10% of the births), and one and a half times as many births as Sunday (12% of the births).

Thus, we can conclude that the main difference from a uniform

> model for local births is that weekend births are particularly uncommon, while Tuesday appears to have the highest frequency of births.

Looking at the details of the χ^2 *test is appropriate only when the test has been shown to be statistically significant.* If not, any deviation from the model assumed in H_0 could very well be the kind of random variations we would expect to see when H_0 is true. That is, differences between chi-square components or differences between sample proportions are meaningless when the overall χ^2 test is not significant.

CAUTION

APPLY YOUR KNOWLEDGE

21.7 Saving birds from windows, continued. Exercise 21.1 described an experiment designed to figure out whether tilting windows down so that they reflect earth rather than sky could help reduce accidental bird strikes. Refer to your calculations in Exercises 21.3 and 21.5 to fully interpret the result of the chi-square test and draw a conclusion about the efficacy of tilting windows.

CONDITIONS FOR THE CHI-SQUARE TEST

The chi-square test for goodness of fit is a one-sample test for categorical data. As with all the statistical tests we have studied, this test is valid only when certain conditions are met. First, the data must satisfy a *multinomial setting*.

THE MULTINOMIAL SETTING

1. There is a fixed number n of observations.
2. The n observations are all **independent**. That is, knowing the result of one observation does not change the probabilities we assign to other observations.
3. Each observation falls into just one of a finite number k of complementary and mutually exclusive outcomes.
4. The probability of a given outcome is the same for each observation.

If you studied optional Chapter 12, you will notice the similarity to the binomial setting. In fact, the binomial setting is a specific case of the multinomial setting.

The second set of requirements for a valid test is related to issues of sample size. The chi-square test, like the z procedures, is an approximate method that becomes more accurate as the counts for each outcome get larger. We must therefore check that the counts are large enough to allow us to trust the *P*-value. Fortunately, the chi-square approximation is accurate for quite modest counts. Here is a practical guideline.[5]

EXPECTED COUNTS REQUIREMENT

You can safely use the chi-square test for goodness of fit for the distribution of one categorical variable with k possible outcomes (levels) when *most of the* k **expected counts**—at least 80%, or 4 out of 5—have a value of 5.0 or greater and *all* k **expected counts** are 1.0 or greater.

Note that the *P*-value of the chi-square test is the right-tail area under a chi-square distribution because the X2 X^2 test statistic has a chi-square distribution *when the null hypothesis is true*. This is why the requirement for the chi-square test involves *expected* counts rather than observed counts.

EXAMPLE 21.8 Not on the weekend, please: conditions

We can now complete Example 21.7 by checking the conditions for inference. The data are one random sample of n=700 $n = 700$ births, and each birth falls on one of the 7 days of the week—the 7 possible outcomes for this variable. Births are independent events because knowing which day one particular birth occurred says nothing about the day of another birth. The probability that a birth falls on one particular day is the same for all births.

The expected counts for a null hypothesis of births being evenly distributed over the week appear in the Minitab outputs in Figure 21.3. The smallest expected count is 100, which is clearly larger than 5.0.

Because the outcomes satisfy the multinomial setting and the expected counts are large enough, the data meet the guideline for safe use of the chi-square test.

EXAMPLE 21.9 Righting behavior of aphids: conditions

In Example 21.2 we looked at the outcome of dropping live aphids from tweezers to assess whether aphids are more likely than chance alone would predict to land right side up (on their legs).

The study videotaped a sample of n=20 $n = 20$ live aphids as they were dropped. Each aphid could land only on its legs or on its back, with an outcome independent of what other aphids experienced. All aphids had the same probability of landing on their legs. The data satisfy a multinomial setting (in fact, with only two possible outcomes, a binomial setting).

Based on a null hypothesis of equal probability of landing either on

> the legs or on the back, the expected counts are both 10 (as shown in the Minitab output displayed in Example 21.6), so both are greater than 5.0.
>
> Therefore we can safely use the chi-square goodness-of-fit test for these data.

The fact that only 1 aphid landed on its back is not relevant to the validity of the chi-square test. The count requirement for the chi-square test applies only to the expected counts. There are no constraints on the value of each observed count.

APPLY YOUR KNOWLEDGE

21.8 Genetics of seed color, continued. Go back to Example 21.1, examining the possibility of a dominant epistatic genetic basis for the color distribution of amaranth grains. Explain why a chi-square procedure is appropriate.

21.9 Saving birds from windows, continued. Go back to Exercise 21.1, describing an experiment with tilted windows to help reduce accidental bird strikes. Explain why a chi-square procedure is appropriate.

21.10 Heat resistance in rice, continued. Go back to Exercise 21.2 investigating the genetic model of heat resistance in rice plants. Explain why a chi-square procedure is appropriate.

21.11 Are the conditions met? You want to run a chi-square goodness-of-fit test on multinomial data from one random sample.

a. Based on the following observed and expected counts, are the conditions met for this test? Explain your answer.

Observed	3	8	9	4
Expected	6	6	6	6

b. Based on the following observed and expected counts, are the conditions met for this test? Explain your answer.

Observed	7	9	7	13
Expected	4	8	8	16

P&R Fotos/age fotostock/SuperStock

CHAPTER 21 SUMMARY

- The **chi-square test for goodness of fit** tests the null hypothesis H_0 that the data come from a population with a given distribution. It specifies a proportion for each of the k possible outcomes in the distribution, and the sum of all these proportions equals 1. The alternative hypothesis H_a simply says that H_0 is not true.
- The test compares the observed counts of observations with the counts that would be expected if H_0 was true. With n total observations, the **expected count** for any given outcome i is

  expected counti=n×pi0

  $$\text{expected count}_i = n \times p_{i_0}$$

  The **chi-square statistic** is a sum of components computed separately for each of the k possible outcomes in the distribution:

  X2=∑(observed count-expected count)2expected count

  $$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

  The *P*-value of the chi-square test for goodness of fit is the area to the right of the test statistic X^2, under the **chi-square distribution** with $k - 1$ **degrees of freedom**. Use this *P*-value to decide whether to reject or fail to reject H_0.
- The chi-square distribution is an approximation to the distribution of the statistic X^2. You can safely use this approximation when all *expected counts* are at least equal to 1.0 and no more than 20% of them are less than 5.0.
- When the chi-square test finds a statistically significant departure from the hypothesized distribution, perform data analysis to describe the nature of the distribution. You can do this by comparing the percents of each outcome, comparing the observed counts with the expected counts, and looking for the largest **components of the chi-square statistic**.

THIS CHAPTER IN CONTEXT

In Chapter 19 we computed confidence intervals and tested hypotheses about the proportion pp of successes in one population based on data from one SRS from this population. That is, for each individual in the sample, we had a binary categorical variable of a success/failure type. This variable could be, for example, a patient's survival (yes/no) or a person's gender (female/male). For more complex categorical variables, such as a survey respondent's opinion (support/oppose/not sure), we selected one outcome as success (support) and treated all other outcomes as failure (not support).

The chi-square method introduced in this chapter lets us deal with categorical variables in a more nuanced way, by considering all levels of the variable. This is particularly relevant for genetics testing when we examine a trait expressed in more than just two phenotypes (for example, black, brown, and pale seeds). The chi-square goodness-of-fit approach compares the counts of each outcome observed in the sample with the hypothetical counts we would expect for these outcomes if our null hypothesis was true.

The null hypothesis of the chi-square test for goodness of fit represents how you believe (or wonder if) the population of interest breaks down into kk levels of the categorical variable. This is similar to the concept of a pie chart for a categorical variable, which we discussed in Chapter 1. Under the null hypothesis, the pie chart depicts the entire population, and the sizes of the slices correspond to the specified proportions for each level (p1,p2,…,pk)$(p_1, p_2, \ldots, p_k)$ of the variable. This explains why you must be sure that these hypothesized proportions all add up to 1 (or 100%, the entirety of the population).

In Chapter 22 we will see how the chi-square statistic can be used to test hypotheses about the relationship between two categorical variables—a comprehensive expansion of the zz test for comparing two population proportions described in Chapter 20.

CHECK YOUR SKILLS

Example 21.1 described a model of dominant epistasis with a 12:3:1 phenotypic ratio in the second generation (F2). A cross of white and green summer squash plants gives the following numbers of squash in F2: 131 white squash, 34 yellow squash, and 10 green squash. Are these data consistent with a 12:3:1 dominant epistatic model of genetic inheritance (white being dominant)? Use this information for Exercises 21.12 through 21.17.

21.12 What is the null hypothesis for the corresponding chi-square goodness-of-fit test?

a. H0:H_0 : The distribution is not uniform
b. H0:H_0 : The distribution is uniform: pwhite=pyellow=pgreen $p_{white} = p_{yellow} = p_{green}$
c. H0:H_0 :pwhite=12/16$p_{white} = 12/16$ and pyellow=3/16$p_{yellow} = 3/16$ and pgreen=1/16$p_{green} = 1/16$

21.13 What is the alternative hypothesis?

a. Ha:H_a : The distribution is not uniform
b. Ha:H_a :pwhite≠pyellow≠pgreen$p_{white} \neq p_{yellow} \neq p_{green}$
c. Ha:H_a :H0H_0 is not true

21.14 What is, approximately, the expected count of green squash?

a. 10.9
b. 1.0
c. 0.06

21.15 What is the value of the chi-square statistic for this test?

a. X2=0.080$X^2 = 0.080$
b. X2=0.124$X^2 = 0.124$
c. X2=0.940$X^2 = 0.940$

21.16 Obtain the *P*-value for this test. What can you conclude?

a. The findings prove that squash color is determined by dominant epistasis.
b. The findings are consistent with the ratio of squash colors determined by dominant epistasis.
c. The findings suggest that squash color is not determined by dominant epistasis.

21.17 You can trust the validity of this test because

a. all expected counts are larger than 5.
b. all observed counts are larger than 5.
c. the sample size is larger than 40.

A very large pediatrics office is open for business Monday through Friday. Management wants to know if antibiotics are prescribed uniformly over the workweek (Monday through Friday). A random sample of 200 antibiotic prescriptions written over the past 12 months is taken. Use this information for Exercises 21.18 and 21.19.

21.18 What is the null hypothesis for the chi-square goodness-of-fit test?

a. H0:H_0 : The distribution is not uniform
b. H0:H_0 : The distribution is uniform: pMon=pTue=pWed=pThu=pFri=1/5 $p_{Mon} = p_{Tue} = p_{Wed} = p_{Thu} = p_{Fri} = 1/5$
c. H0:H_0 :p=05$p = 05$.

21.19 Software gives P<0.01$P < 0.01$. Your conclusion based on this test is that

a. the data are consistent with a uniform distribution.
b. the data provide strong evidence that all five proportions are different.
c. the data provide strong evidence that the distribution isnot uniform.

Example 21.1 examined the distribution of amaranth seed color. Another genetic model that could explain the existence of exactly three phenotypes *is based on a single-gene mode of inheritance with two codominant alleles. Under this model, crossing pure breeds of the dominant and recessive traits would give rise to the two original traits as well as an intermediate phenotype in F2, with an expected 1:2:1 ratio of black, brown, and pale seeds, respectively. A genetic crossing resulted in 321 black seeds, 77 brown seeds, and 31 pale seeds. Use this information for Exercises 21.20 and 21.21.*

21.20 What is the null hypothesis for the corresponding chi-square goodness-of-fit test?

a. H0:H_0 :pblack=pbrown=ppale=1/3$p_{black} = p_{brown} = p_{pale} = 1/3$

b. H0:H_0 :pblack=0.25$p_{black} = 0.25$ and pbrown=0.5$p_{brown} = 0.5$ and ppale=0.25$p_{pale} = 0.25$

c. H0:H_0 :pblack=0.75$p_{black} = 0.75$ and pbrown=0.18$p_{brown} = 0.18$ and ppale=0.07$p_{pale} = 0.07$

21.21 The R output for the appropriate chi-square goodness-of-fit test is X-squared = 568.3566, df = 2, p-value < 2.2e-16. What can you conclude about amaranth seed color?

a. The findings prove that amaranth seed color isdetermined by one gene with two codominant alleles.

b. The findings are consistent with a color ratio for amaranth seeds determined by one gene with two codominant alleles.

c. The findings provide very strong evidence that amaranth seed color isn’t determined by one gene with two codominant alleles.

CHAPTER 21 EXERCISES

21.22 Mendel's first law. Gregor Mendel has been called the "father of genetics," having formulated laws of inheritance even before the function of DNA was discovered. In one famous experiment with peas, he crossed pure breeds of tall and dwarf pea plants. The first-generation pea plants were all tall. In the second generation (F2), he obtained 787 tall plants and 277 dwarf plants. Mendel's first law of inheritance is a model with a dominant and a recessive phenotype. The first law says that we can expect a 3:1 ratio—that is, 3/4 with the dominant phenotype and 1/4 with the recessive phenotype—in F2.

a. Are the conditions for a chi-square goodness-of-fit test satisfied?
b. What are the null and alternative hypotheses for this test?
c. What are the expected counts under the null hypothesis?
d. Obtain the chi-square statistic and the test *P*-value. What can you conclude? Do these data support Mendel's first law?

21.23 Mendel's second law. Mendel formulated the law of independent assortment as his second law of inheritance. It involves two genes independently segregated during reproduction, each independently determining one aspect of the phenotype. In one experiment, Mendel crossed pea plants producing yellow, round seeds with pea plants producing green, wrinkled seeds. The first generation resulted in only plants producing yellow, round seeds. Self-crossing of the F1 plants yielded the following phenotypes in F2:

Phenotype	Yellow, round	Yellow, wrinkled	Green, round	Green, wrinkled
Counts	315	101	108	32

Assuming two independent genes, each with a dominant allele and a recessive allele, we would expect to find a 9:3:3:1 phenotypic ratio.

a. Are the conditions for a chi-square goodness-of-fit test satisfied?
b. What are the null and alternative hypotheses for this test?
c. What are the expected counts under the null hypothesis?
d. The chi-square statistic for this test is X2=0.47$X^2 = 0.47$. Obtain the test *P*-value. What can you conclude? Do the data agree with Mendel’s second law of genetic inheritance?

21.24 Nematode resistance in soybeans. The soybean cyst nematode is a roundworm pest affecting the yield of soybean crops worldwide. Nematode-resistant soybean lines have been used for agriculture. One recent soybean line we will call A is resistant to most nematodes except the supervirulent nematode LY1. A new soybean line we will call B has been shown to be resistant to LY1 and to be genetically different from soybean line A. Understanding the genetics of soybean resistance to LY1 is an important step in creating a genetically engineered soybean that will resist most nematodes, including the virulent LY1. Researchers crossed soybean lines A and B. They obtained 105 second-generation plants, of which 5 were resistant to LY1 and 100 were not. The researchers state that the data are “a good fit to an expected three-gene model, Rhg, rhg, rhg (one dominant and two recessive genes) with a segregation ratio of 3 (Resistant):61 (Susceptible).”[6]

a. The researchers describe a 3:61 distribution. What are the null and alternative hypotheses for the corresponding chi-square goodness-of-fit test?
b. What are the expected counts under the null hypothesis? Are the conditions for the test satisfied?
c. Obtain the chi-square statistic and the *P*-value for this test. Do the data support the researchers’ proposed model?
d. Can you conclude that the genetic basis of soybean resistance to the virulent LY1 nematode is carried by three genes, one dominant and two recessive? Explain your answer.

21.25 Plan B emergency contraception. Plan B is the common name in the United States and Canada for the emergency contraception drug levonorgestrel. The high dose of progestogen it delivers has been shown in the laboratory to inhibit ovulation, thus rendering fertilization impossible. Some groups, however, are concerned that if the drug interfered with the fertilized egg, it could be considered abortive rather than strictly contraceptive. Researchers enrolled a cohort of fertile women attending a family-planning clinic for emergency contraception after unprotected intercourse during their fertile period. From interviews, blood samples, and ultrasounds, the researchers were able to establish the time of intercourse relative to the women’s menstrual cycles. In this cohort, 87 women had unprotected intercourse before ovulation (pre-ovulation group) and 35 women had unprotected intercourse on or after ovulation (post-ovulation group). Based on standard clinical assessments of fertility rates by day of the menstrual cycle, the researchers expected 13.2 pregnancies in the pre-ovulation group and 7.1 pregnancies in the post-ovulation group. Instead, they observed 0 and 6 pregnancies, respectively.[7]

a. We want to know if there is evidence that Plan B impacts the chance of pregnancy when taken before ovulation. What are the observed and expected counts of pregnancy and no pregnancy in the pre-ovulation group? Use these counts to compute the chi-square statistic. Are the test assumptions met? Obtain the test *P*-value and conclude.

b. We want to know if there is evidence that Plan B impacts the chance of pregnancy when taken after ovulation. What are the observed and expected counts of pregnancy and no pregnancy in the post-ovulation group? Use these counts to compute the chi-square statistic. Are the test assumptions met? Obtain the test *P*-value and conclude.

c. Write a short paragraph describing the study findings and what they suggest about the mechanism of action of the Plan B drug. Include the test statistics and *P*-values in your report.

21.26 Octopus locomotion. With eight arms arranged in a circular pattern, octopuses can crawl in any direction. This crawling locomotion is accomplished by pushing against a surface with one of the eight arms. Researchers examined a random sample of 2506 nonconsecutive short video clips of a moving octopus and found the distribution of pushing arms shown in Figure 21.4. The published report includes the following statement:[8] “For the hypothesis that all the eight arms usage frequency was uniformly distributed: df=7,X2=163.478,p<0.0001$df = 7, X^2 = 163.478, p < 0.0001$.”

a. State the null and alternative hypotheses the researchers tested.
b. Verify that the test assumptions are met.
c. Use Figure 21.4 and the cited chi-square analysis to write a short paragraph describing the study findings and what they suggest about crawling locomotion in the octopus.

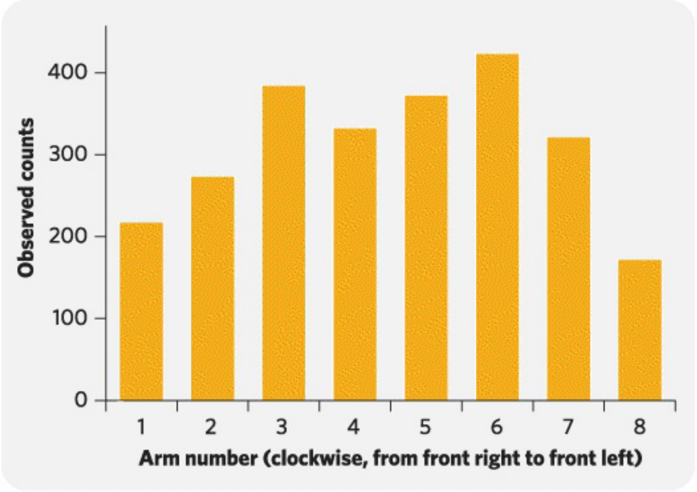


Figure 21.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 21.4 Distribution of pushing arm in a moving octopus. The numbers 1 through 8 refer to the clockwise location of the arm on the body, starting with the front right location (arm 1).

21.27 America’s favorite alcoholic beverage. Are beer, wine, and liquor equally favored by Americans as the alcoholic drink they consume most often? A 2016 Gallup survey interviewed a random sample of American adults about their alcohol consumption and preferences. Of the 647 non-abstainers who answered the question about which beverage—beer, wine, or liquor—they consume most often, 293 selected beer, 218 selected wine, and the remaining 136 selected liquor. Follow the four-step process as illustrated in Example 21.7.

21.28 Characteristics of a forest. Forests are complex, evolving ecosystems. For instance, pioneer tree species can be displaced by successional species better adapted to the changing environment. Ecologists mapped a large Canadian forest plot dominated by Douglas fir with an understory of western hemlock and western red cedar. Sapling trees (young trees shorter than 1.3 meters) are indicative of the future of a forest. The 246 sapling trees recorded in this sample forest plot were of the following types:[9]

Tree species	**Count of sapling trees**
Western red cedar (RC)	164
Douglas fir (DF)	0
Western hemlock (WH)	82

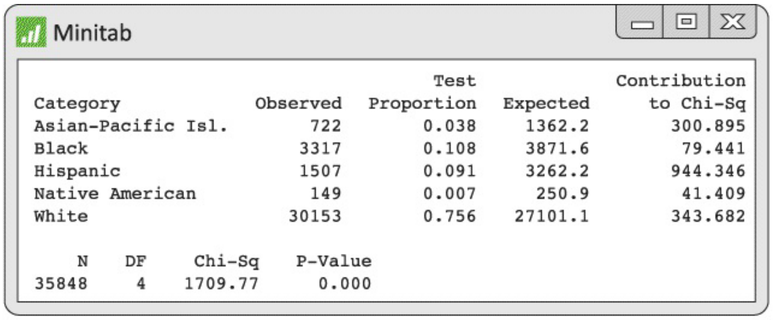
Are there equal proportions of RC, DF, and WH among sapling trees in this forest? What does your analysis suggest about this forest's successional stage? Follow the four-step process as illustrated in Example 21.7.

21.29 Are clinical trials representative of the American population? Individuals of different ethnicities are somewhat more genetically variable than those of the same ethnicity and may respond to treatments differently. The FDA has recently requested that every effort be made to include individuals of all major ethnic backgrounds when enrolling subjects in clinical trials. A randomly chosen article in the *New England Journal of Medicine* describes the efficacy of calcium and vitamin D supplementation in preventing hip and other fractures in healthy postmenopausal women. Here is the breakdown by major ethnic group of the subjects whose ethnicity is known, along with the percent of individuals from each ethnicity in the general American population:[10]

Ethnicity	**Count in study**	**Percent in U.S.**
White	30,153	75.6

Black	3,317	10.8
Hispanic	1,507	9.1
Asian or Pacific Islander	722	3.8
Native American	149	0.7

Did this study successfully match the ethnic diversity of the American population? Figure 21.5 shows the Minitab output for this test. What does your analysis suggest? Follow the four-step process.



```
                               Test               Contribution
Category          Observed  Proportion  Expected    to Chi-Sq
Asian-Pacific Isl.     722       0.038    1362.2      300.895
Black                 3317       0.108    3871.6       79.441
Hispanic              1507       0.091    3262.2      944.346
Native American        149       0.007     250.9       41.409
White                30153       0.756   27101.1      343.682

    N   DF    Chi-Sq   P-Value
35848    4   1709.77     0.000
```

Figure 21.5

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 21.5 Minitab output for Exercise 21.29.

21.30 Weight perception. A Gallup survey asked 1014 adults aged 18 and older about their ideal weight. The survey found that 116 interviewees thought that they were currently under their ideal weight, 180 thought that they were at about their ideal weight, and 718 thought that they were over their ideal weight. Do these self-perceptions match the reality of the weight distribution in the United States? The National Health Interview Survey estimates, based on people’s actual body mass index (computed from both weight and height), that 1.8% of the U.S. adult population is underweight, 36.7% has a healthy weight, and 61.5% is either overweight or obese. Use these official percents to compute the expected counts. Follow the four-step process. (Software gives X2=636.9$\chi^2 = 636.9$.)

21.31 Inbreeding in prairie dogs. A field biologist recorded events of inbreeding in a colony of prairie dogs. Based on a sample of 44 estrous females and 17 sexually mature males, the researcher computed the coefficient of relatedness of all possible male–female pairs in this sample. On the one hand, if individuals avoided inbreeding, we would expect to observe breeding mostly between individuals with little genetic relatedness. On the other hand, if breeding occurred randomly, without regard for the genetic relatedness of the mating couple, we would expect breeding observations among individuals of various relatedness to be proportional to the number of possible pairs with a given coefficient of relatedness. Here is the table of observed and expected breeding pairs in thisstudy:[11]

Coefficient of relatedness	**Observed pairs**	**Expected pairs**
$r\geq1/4$ (siblings)$r \geq 1/4\ (\text{siblings})$	6	6
$1/4>r\geq1/8$$1/4 > r \geq 1/8$	4	5
$1/8>r\geq1/16$$1/8 > r \geq 1/16$	18	10
$1/16>r\geq1/32$$1/16 > r \geq 1/32$	12	12
$1/32>r\geq1/64$$1/32 > r \geq 1/64$	10	11
$1/64>r\geq1/128$$1/64 > r \geq 1/128$	3	7
$1/128>r\geq0$$1/128 > r \geq 0$	8	8
No known kinship	8	10

Software computes $X2=9.377$$X^2 = 9.377$. Do the data support the hypothesis that breeding among prairie dogs occurs randomly, without regard for genetic relatedness? Follow the four-step process.

21.32 Genetics of plant variegation. *Arabidopsis thaliana* is a green-leafed plant, but a variegated type has green-and-yellow leaves. To investigate the mechanism of plant variegation, researchers created, in the DNA of the variegated *Arabidopsis*, a recessive mutation that suppresses variegation.[12] When the green mutants were crossed with the variegated types, the first-generation plants were all variegated, confirming the recessive aspect of the mutation. In the second generation (F2), the researchers obtained 84 green plants and 216 variegated plants. Are these data consistent with a single recessive mutation for which we would expect a 3:1 ratio in F2? Follow the four-step process.

21.33 Genetics of plant variegation, continued. The researchers in Exercise 21.32 also crossed mutant green plants with wild-type green plants. Not surprisingly, the first generation was 100% green. In the second generation (F2), however, a new phenotype was observed in some of the plants: a transitional variegation occurring only on the first few leaves. (We will call them “transitional.”) Of the 85 plants obtained in F2, 65 were wild-type green, 15 variegated, and 5 transitional. The researchers hypothesized that these phenotypes are the result of two genes following a dominant epistatic model. Are these data consistent with a dominant epistatic model for which we would expect a 12:3:1 distribution in F2? Follow the four-step process.

21.34 Transgenic chickens. Infection of chickens with the avian flu is a threat to both poultry production and human health. A research team created transgenic chicken resistant to avian flu infection. If the transgene is inherited normally without lethal effects, then transgenic and wild-type crosses should provide, in the second generation (F2), 25% homozygote transgenic offspring, 50% heterozygote offspring, and 25% homozygote wild-type offspring. The researchers obtained 102 chickens in F2, 25 of which were homozygote transgenic, 45 heterozygote, and 32 homozygote wild-type.[13] Are these data consistent with a 1:2:1 ratio in F2 as expected from a transgene inherited normally without lethal effects? Follow the four-step process.

21.35 Frizzled feathers. The Frizzle fowl is a striking variety of chicken with curled feathers. In a 1930 experiment, Frizzle fowls were crossed with a Leghorn variety exhibiting straight feathers. The first generation (F1) produced all slightly frizzled chicks. When the F1 was interbred, the following characteristics were observed in F2:[14]

Phenotype (feather type)	**Observed counts**
Frizzled	23
Slightly frizzled	50
Straight	20

The most likely genetic model for these results is that of a single gene locus with two codominant alleles. Under such a model, we would expect a 1:2:1 ratio in F2. Do the data support a codominance model of inheritance for this feather phenotype? Follow the four-step process.

21.36 Poisson distribution? (optional) As part of a class project, students dropped split peas onto a surface marked with a 20 cm by 20 cm grid made up of 100 equal-sized quadrats (each 4 cm 2^{2} in surface area). The peas were dropped blindly from a height of one foot above the grid. The students then counted how many peas had landed on each of the 100 quadrats. Here are the results in summarized form:[15]

Number of peas per quadrat	0	1	2	3	≥4≥ 4
Number of quadrats	45	36	16	3	0

a. Verify that the number of quadrats adds up to 100 and that the number of peas that landed on the grid is 77. What is the average number of peas per quadrat?

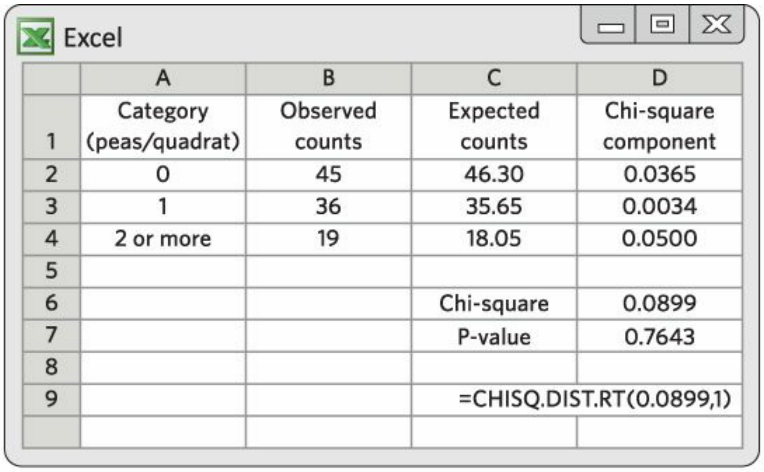
b. We described the Poisson probability distribution in Chapter 12. You might want to refresh your memory. If the throwing of the split peas was truly random, we would expect the number of peas per quadrat to follow a Poisson distribution with mean equal to the mean number of peas per quadrat. The value you computed in part a based on the sample data is your best guess for the true value of the Poisson parameter. Obtain P(X=k) $P(X = k)$ for XX equals 0, 1, 2, 3, and 4 or more using this Poisson distribution.

c. We now want to test whether there is significant evidence that the distribution of split peas on the grid is not Poisson. For that purpose, we can use the chi-square test for goodness of fit. Use the values you computed in part b to obtain the expected counts. Are the conditions for inference satisfied?

d. We need to group the data so that the expected counts are not too low. We can pool the results for 2, 3, and 4 or more peas per quadrat into a single category “2 or more peas per quadrat.” What is P(X≥2)?$P(X \geq 2)$? Verify that the conditions for inference are now satisfied.

e. Figure 21.6 shows the details of the chi-square test performed with Excel. Verify that you obtain the same value for the chi-square statistic on your own. Notice that we used only 1 degree of freedom to find the *P*-value, because we lost 1 degree of freedom when we used sample data to estimate the true population parameter μμ of the Poisson distribution.

f. Does the test provide significant evidence that the distribution of split peas on the grid is not Poisson? That is, is there significant evidence that the throwing was not completely random?



	A	B	C	D
1	Category (peas/quadrat)	Observed counts	Expected counts	Chi-square component
2	0	45	46.30	0.0365
3	1	36	35.65	0.0034
4	2 or more	19	18.05	0.0500
5				
6			Chi-square	0.0899
7			P-value	0.7643
8				
9			=CHISQ.DIST.RT(0.0899,1)	

Figure 21.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 21.6 Excel calculations for Exercise 21.36.

CHAPTER 22 The Chi-Square Test for Two-Way Tables

DCPhoto/Alamy Stock Photo

IN THIS CHAPTER WE COVER...

- Two-way tables
- Hypotheses for two-way tables of counts
- Expected counts and chi-square statistic
- The chi-square test
- Conditions for the chi-square test
- The chi-square test and the two-sample z test

In Chapter 21 we introduced a new statistic, X2,χ^2, and used it to test the null hypothesis that a categorical variable follows (or “fits”) a particular distribution. This is a situation often found in genetic analysis. However, the most common use of the chi-square statistic is to test the null hypothesis that there is *no relationship between two categorical variables.*

Some variables—such as sex, race, and blood type—are categorical by nature. Other categorical variables are created by grouping values of a quantitative variable into classes—such as age groups (for example, 20–29, 30–39, and so on) and weight categories (for example, underweight, healthy, overweight, obese). As usual, we are interested in both data analysis (describe the relationship) and inference (Is the relationship statistically significant?).

TWO-WAY TABLES

Chapter 5 described how to present data on two categorical variables in a **two-way table** of counts. That's our starting point. Here is an example.

EXAMPLE 22.1 Hormone therapy and thromboembolism

Oral estrogen-replacement therapy (ERT) activates blood coagulation and increases the risk of venous thromboembolism (a traveling blood clot, sometimes fatal) in postmenopausal women. Transdermal ERT might be less problematic, at least in theory. A retrospective, case-control study compared a random sample of 155 postmenopausal women with a first documented thromboembolism and a control random sample of 381 postmenopausal women with no history of thromboembolism. The women were asked about their ERT history. Here are the results:[1]

	Thromboembolism status	
ERT history	**Cases**	**Controls**
Never use	71	208
Past use	22	53
Current oral ERT use	32	27
Current transdermal ERT use	30	93
Total	155	381

Getty Images

The two-way table in Example 22.1 shows the relationship between *two* categorical variables. One variable is the patient's ERT history, with four levels. The other variable is the patient's thromboembolism status: first thromboembolism (cases) or no thromboembolism (controls). The two-way table gives the counts for all eight combinations of values of these variables. Each of the eight counts occupies a **cell** of the table.

It is difficult to compare the counts because the control sample is much larger than the case sample. Here are the percents of each outcome within each sample:

	Thromboembolism status	
ERT history	**Cases**	**Controls**
Never use	46%	55%
Past use	14%	14%
Current oral ERT use	21%	7%
Current transdermal ERT use	19%	24%
Total	100%	100%

In the language of Chapter 5 (page 127), these are the **conditional distributions** of ERT history, given the patients' thromboembolism status. The differences are not large, but the case group has a higher percent of women who are current oral ERT users. Figure 22.1 compares the two distributions. We want to know if there is a significant difference between the two distributions of ERT history.

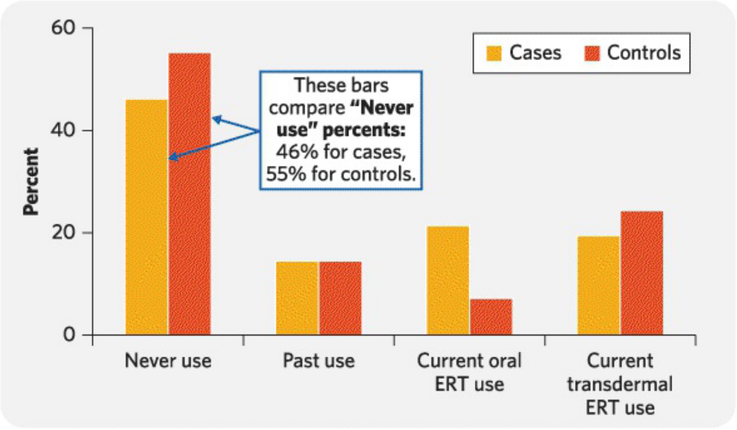


Figure 22.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.1 Bar graph comparing hormone therapy (ERT) history for postmenopausal

women with (cases) or without (controls) a first diagnosis of thromboembolism.

APPLY YOUR KNOWLEDGE

22.1 E-cigarettes and smoking cessation. Electronic cigarettes (e-cigarettes) are battery-operated devices that can deliver nicotine and other chemicals in water vapor instead of smoke. They are a relatively recent invention with little scientific evaluation, yet some people believe that e-cigarettes could help smokers who want to quit smoking. Researchers in New Zealand randomly assigned 657 adult smokers wanting to quit to three treatment groups: nicotine patches, nicotine e-cigarettes, and placebo e-cigarettes. Smoking abstinence at the end of the study was assessed biochemically for all participants with a breath analyzer. The published research reports that, "at 6 months, verified abstinence was 7.3% (21 of 289) with nicotine e-cigarettes, 5.8% (17 of 295) with patches, and 4.1% (three of 73) with placebo e-cigarettes."[2]

a. Organize the study findings in the following two-way table of counts. Add the group totals to your table.

	Smoking status at six months		
Treatment group	**Verified abstinence**	**Non-abstinent**	**Total**
Nicotine e-cigarettes			
Patches			
Placebo e-cigarettes			

b. Now organize the cited percents into a similar two-way table. Are these percents part of the conditional distribution of abstinence, given treatment group, or are they part of the conditional distribution of treatment, given abstinence status? What do they suggest about the relationship between treatment assigned and abstinence outcome?

22.2 Aphid righting behavior. Pea aphids, *Acyrthosiphon pisum*, are small, wingless, sap-sucking insects that live on plants. They evade predators (such as ladybugs) by dropping off the plants. A study examined the mechanism of aphid drops. Researchers placed aphids on a leaf positioned at four different heights (in centimeters, cm) above a surface covered with petroleum jelly. When a ladybug was introduced, the aphids dropped, and the petroleum jelly helped capture their landing posture (upright or not). Each aphid performed this experiment only once. Here are the findings:[3]

	Dropping height			
Landing posture	**3 cm**	**5 cm**	**10 cm**	**20 cm**
Upright	20	23	27	29
Not upright	10	7	3	1
Sample size	30	30	30	30

a. What percent of aphids land upright when dropping from 3 cm? From 5 cm? From 10 cm? From 20 cm? These percents do not add up to 100% because these groups represent separate samples.

b. Make a bar graph that compares your percents for all four samples. Is there any clear relationship between dropping height and landing posture? Explain your answer.

Perennou Nuridsany/Science Source

HYPOTHESES FOR TWO-WAY TABLES OF COUNTS

In this chapter we examine the potential pattern of association between two categorical variables. As with the other inference procedures we have seen, the null hypothesis represents a statement of “no effect.” Therefore, the null hypothesis states that there is *no association (no relationship)* between these two categorical variables at the population level. In the probabilistic language of Chapter 10, this also means that the two categorical variables are **independent.**

EXAMPLE 22.2 Hormone therapy and thromboembolism: hypotheses

In Example 22.1 we compared the distribution of ERT history among postmenopausal women with or without a first diagnosis of thromboembolism in a case-control study. The null hypothesis is that there is *no difference* between the distributions of ERT history in the corresponding two populations. Because ERT history and thromboembolism status form two categorical variables, the null hypothesis can also state that there is *no relationship* between these two categorical variables:

> H0:H_0 : There is no relationship between thromboembolism status and ERT history in the corresponding populations.

The alternative hypothesis says that there *is* a relationship but does not specify any particular kind of relationship:

> Ha:H_a : There is some relationship between thromboembolism status and ERT history in the corresponding populations. (H0H_0 is not true.)

The alternative hypothesis is not one-sided or two-sided. It is nonspecific and allows any kind of difference. Any significant difference between these distributions indicates that the null hypothesis is false and the alternative hypothesis is true.

Using the zz procedures of Chapter 20, we might start by comparing the proportions of patients in the two groups who have never used ERT. We could similarly compare the proportions for each of the other ERT histories: four tests in all, with four *P*-values. This is a bad idea. *When we obtain many individual tests or confidence intervals, the individual P-values and confidence levels don't tell us how confident we can be in all of the inferences taken together*.

CAUTION

For this reason, it's also cheating to pick out the largest of the four differences and then test its significance as if it were the only comparison we had in mind. For example, the "current oral ERT use" proportions in Example 22.1 are significantly different (P<0.0001)$(P < 0.0001)$ if we compare just this one outcome. But is it surprising that the *most different* proportions among four ERT history levels differ by this much? That's a different question.

The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions is common in statistics. This is the problem of **multiple comparisons.** Statistical methods for dealing with multiple comparisons usually have two steps:

1. An *overall test* to see whether there is good evidence of *any* differences among the parameters that we want to compare (using a nondirectional alternative hypothesis).
2. A detailed *follow-up analysis* to decide which of the parameters differ and to estimate how large the differences are.

The overall test, though more complex than the tests we saw earlier, is often reasonably straightforward. The follow-up analysis, however, can be quite elaborate. In our basic introduction to statistical practice, we will concentrate on the overall test, along with descriptive data analysis that points to the nature of the differences.

APPLY YOUR KNOWLEDGE

22.3 E-cigarettes and smoking cessation, continued. Exercise 22.1 describes an experiment comparing three approaches for quitting smoking.

a. State the null and alternative hypotheses that are appropriate for the study’s objective.

b. If we wanted to compare the treatments two at a time, how many two-sample z tests (described in Chapter 20) would be required? Explain why we should run one overall hypothesis test rather than all these two-sample tests.

22.4 Aphid righting behavior, continued. Exercise 22.2 describes a study of aphids dropping from various heights to escape predators.

a. Using the plus four method from Chapter 19, give four 95% confidence intervals for the percents of aphids that land upright when dropping from 3, 5, 10, and 20 cm.

b. Explain clearly why we are *not* 95% confident that *all four* of these intervals capture their respective population proportions.

c. What are the null and alternative hypotheses that are appropriate to address the study’s objective with just one overall test?

DCPhoto/Alamy Stock Photo

EXPECTED COUNTS AND CHI-SQUARE STATISTIC

Our general null hypothesis H0H_0 is that there is *no relationship* between the two categorical variables that label the rows and columns of a two-way table of counts. That is, the two variables are independent. To test H0,H_0, we compare the observed counts in the table with the *expected counts*, the counts we would expect—except for random variation—if H0H_0 was true. Observed counts that differ notably from the expected counts provide evidence against H0.H_0. It is easy to find the expected counts.

EXPECTED COUNTS

The **expected count** in any cell of a two-way table when H0H_0 is true is

expected count= row total × column totaltable total

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

EXAMPLE 22.3 Hormone therapy and thromboembolism: expected counts

Let's find the expected counts for the thromboembolism study. Here is the two-way table with row and column totals:

	Thromboembolism status		
ERT history	**Cases**	**Controls**	**Total**
Never use	71	208	279
Past use	22	53	75
Current oral ERT use	32	27	59
Current transdermal ERT use	30	93	123
Total	155	381	536

The expected count of women with thrombosis (cases) who have never used ERT is

row 1 total × column 1 totaltable total=(279)(155)536=80.68

$$\frac{\text{row 1 total} \times \text{column 1 total}}{\text{table total}} = \frac{(279)(155)}{536} = 80.68$$

Here is the table of all eight expected counts:

	Thromboembolism status		
ERT history	**Cases**	**Controls**	**Total**
Never use	80.68	198.32	279
Past use	21.69	53.31	75
Current oral ERT use	17.06	41.94	59
Current transdermal ERT use	35.57	87.43	123
Total	155	381	536

As this table shows, *the expected counts have exactly the same row and column totals (aside from roundoff error) as the observed counts.*

To see how the data diverge from the null hypothesis, compare the observed counts with these expected counts. You see, for example, that 208 women with no thromboembolism diagnosis (controls) reported never having used ERT, whereas we would expect only 198.32 women to be never-users if the null hypothesis was true. Most notably, 32 women with a first thromboembolism diagnosis (cases) reported currently using oral ERT, which is nearly twice as many women as we would have expected (17.06) if the two variables were independent.

Why the formula works. Where does the formula for an expected cell count come from? The expected count, in n independent trials, for a particular outcome with probability p of occurring is simply np. (Chapter 12 explains why in detail.) This doesn’t mean that we expect to see exactly np counts of that outcome in n trials. Rather, we expect np counts on average over the long run if we repeat the n trials many, many times.

In the thromboembolism study, the proportion of all 536 study participants who reported never having used ERT is

count of never usetable total=row 1 totaltable total=279536

$$\frac{\text{count of never use}}{\text{table total}} = \frac{\text{row 1 total}}{\text{table total}} = \frac{279}{536}$$

Because the study collected sample data, this is our best guess for the value of the parameter p, the overall proportion of postmenopausal women who never used ERT. If H_0 is true, we expect (aside from random variation) the same proportion p of successes in both groups. So the expected count of never use among the 155 case women with a thromboembolism diagnosis is

np=(155)279536=80.68

$$np = (155)\left(\frac{279}{536}\right) = 80.68$$

That's the formula provided in the "Expected Counts" box.

Here is another way of thinking about the expected counts: If H_0 is true and the two variables are independent, then we expect the joint probability of two outcomes to be equal to the product of the corresponding marginal probabilities for each outcome. This is the multiplication rule for independent events described in Chapter 10. In a two-way table, this translates to

expected counttable total=row totaltable total×column totaltable total

$$\frac{\text{expected count}}{\text{table total}} = \frac{\text{row total}}{\text{table total}} \times \frac{\text{column total}}{\text{table total}}$$

Simplifying by "table total" gives the formula in the Expected Counts box.

Expected counts are hypothetical counts and do not have to be whole numbers. However, they must add to the observed row, column, and table totals. That is, the expected counts are simply a statement of how the observations in the study would be expected to fall into the cells of the two-way table, according to H_0.

CAUTION

To compare observed and expected counts in a chi-square test for two-way tables and ask whether there is significant evidence of a relationship between the two categorical variables, we need a test statistic. That test statistic is the *chi-square statistic*, X^2.

CHI-SQUARE STATISTIC

The **chi-square statistic** is a measure of how far the observed counts in a two-way table are from the expected counts defined by the null hypothesis. The formula for the statistic is

X2=∑(observed count-expected count)2expected count

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

The sum is over all cells inside the table. That is, there are as many terms in the sum as there are cells inside the table. Each term in the sum is called a **chi-square component**.

Think of the chi-square statistic X2X^2 as a measure of the distance between the observed counts and the expected counts. Like any distance, it is always zero or positive, and it is zero only when the observed counts are exactly equal to the expected counts. Small values of X2X^2 represent small deviations from H0H_0 that do not provide sufficient evidence to reject H0. H_0. Inversely, large values of X2X^2 are evidence against H0,H_0, because they say that the observed counts are far from what we would expect if H0 H_0 was true.

EXAMPLE 22.4 Hormone therapy and thromboembolism: calculating X2X^2

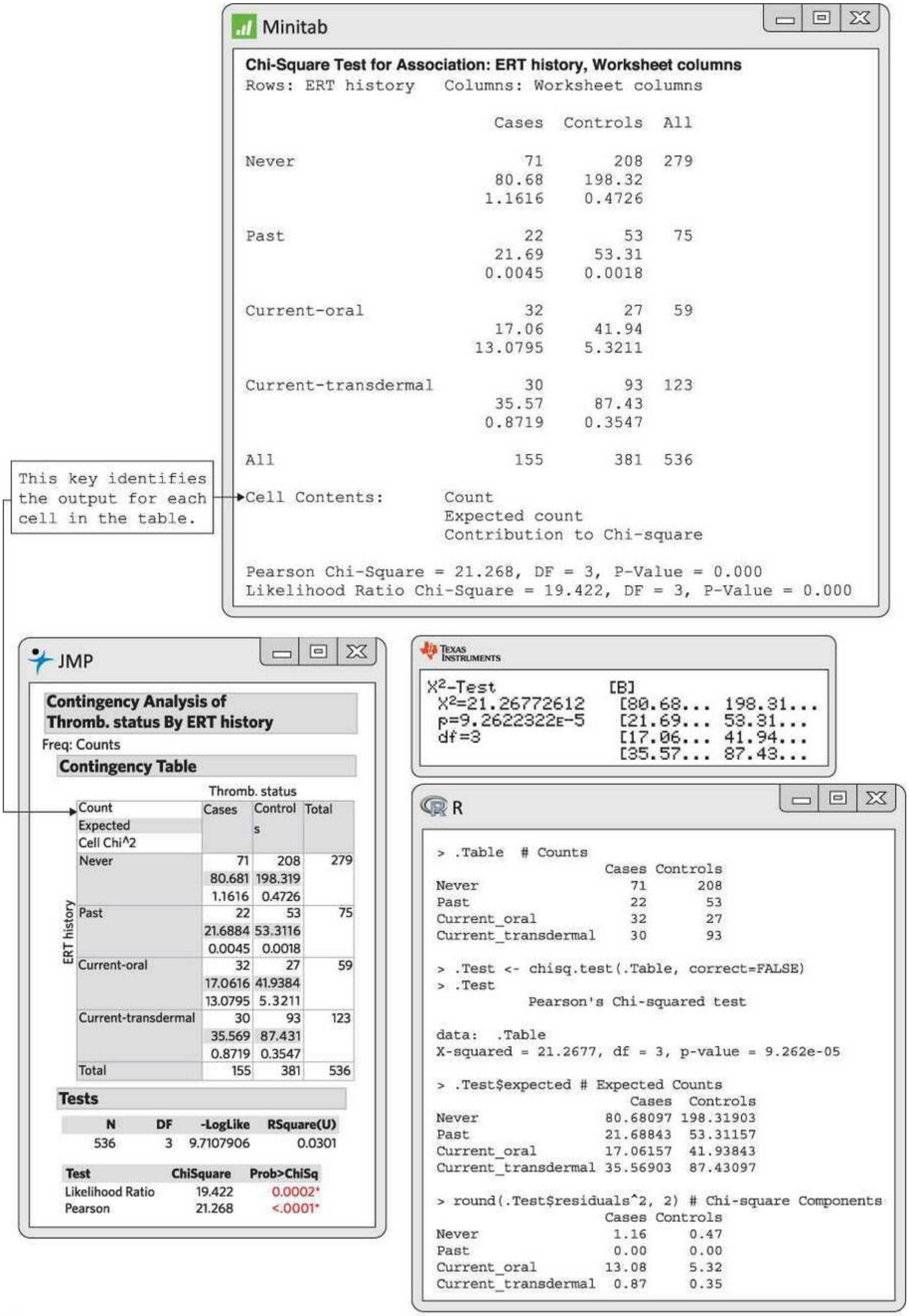
The chi-square statistic is a sum of terms, one for each cell in the table. There are 8 cells in the thromboembolism table of Example 22.1 (2 groups with 4 ERT histories each), so we first need to compute 8 chi-square components separately. Here is how to compute the first one: 71 case women with a thromboembolism diagnosis reported never having used ERT. The expected count for this cell is 80.68. So the chi-square component from this cell is

(observed count - expected count)2expected count=(71-80.68)280.68=93.702480.68=1.16

$$\begin{aligned}\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} &= \frac{(71-80.68)^2}{80.68}\\ &= \frac{93.7024}{80.68} = 1.16\end{aligned}$$

> To obtain the chi-square statistic for this test, we must add all 8 components:
>
> X2=(71-80.68)280.68+(208-198.32)2198.32+(22-21.69)221.69+(53-53.31)253.31+(32-17.06)217.06+(27-41.94)241.94+(30-35.57)235.57+(93-87.43)287.43=21.268
>
> $$\begin{aligned} X^2 &= \frac{(71-80.68)^2}{80.68}+\frac{(208-198.32)^2}{198.32} \\ &\quad +\frac{(22-21.69)^2}{21.69}+\frac{(53-53.31)^2}{53.31} \\ &\quad +\frac{(32-17.06)^2}{17.06}+\frac{(27-41.94)^2}{41.94} \\ &\quad +\frac{(30-35.57)^2}{35.57}+\frac{(93-87.43)^2}{87.43} \\ &= 21.268 \end{aligned}$$

In practice, we let technology handle these computations. Figure 22.2 shows output for the chi-square test for the thromboembolism data from a graphing calculator and three statistical programs. All four display the expected counts and the value of the test statistic, X2=21.268.$X^2 = 21.268.$ The chi-square statistic is a sum of 8 terms, the chi-square components, one for each cell in the table. The three statistical programs provide the details of the chi-square statistic computations, though each uses a different terminology: “Cell Chi^ 2” for JPM, “Contribution to Chi-square” for Minitab, and “Chi-square Components” (or “residuals^ 2”) for R Commander.



Chi-Square Test for Association: ERT history, Worksheet columns

```
Rows: ERT history   Columns: Worksheet columns

                          Cases  Controls  All

Never                        71       208  279
                          80.68    198.32
                         1.1616    0.4726

Past                         22        53   75
                          21.69     53.31
                         0.0045    0.0018

Current-oral                 32        27   59
                          17.06     41.94
                        13.0795    5.3211

Current-transdermal          30        93  123
                          35.57     87.43
                         0.8719    0.3547

All                         155       381  536

Cell Contents:      Count
                    Expected count
                    Contribution to Chi-square

Pearson Chi-Square = 21.268, DF = 3, P-Value = 0.000
Likelihood Ratio Chi-Square = 19.422, DF = 3, P-Value = 0.000
```

This key identifies the output for each cell in the table.

Contingency Analysis of Thromb. status By ERT history

Freq: Counts

Contingency Table

ERT history	Thromb. status Cases	 Controls	 Total
Count Expected Cell Chi^2			
Never	71 80.681 1.1616	208 198.319 0.4726	279
Past	22 21.6884 0.0045	53 53.3116 0.0018	75
Current-oral	32 17.0616 13.0795	27 41.9384 5.3211	59
Current-transdermal	30 35.569 0.8719	93 87.431 0.3547	123
Total	155	381	536

Tests

N	DF	-LogLike	RSquare(U)
536	3	9.7107906	0.0301

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	19.422	0.0002*
Pearson	21.268	<.0001*

```
X²-Test              [B]
 X²=21.26772612       [80.68... 198.31...
 p=9.2622322E-5       [21.69... 53.31...
 df=3                 [17.06... 41.94...
                      [35.57... 87.43...
```

```
> .Table  # Counts
                    Cases Controls
Never                  71      208
Past                   22       53
Current_oral           32       27
Current_transdermal    30       93

> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
              Pearson's Chi-squared test

data:  .Table
X-squared = 21.2677, df = 3, p-value = 9.262e-05

> .Test$expected # Expected Counts
                       Cases  Controls
Never               80.68097 198.31903
Past                21.68843  53.31157
Current_oral        17.06157  41.93843
Current_transdermal 35.56903  87.43097

> round(.Test$residuals^2, 2) # Chi-square Components
                    Cases Controls
Never                1.16     0.47
Past                 0.00     0.00
Current_oral        13.08     5.32
Current_transdermal  0.87     0.35
```

Figure 22.2

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

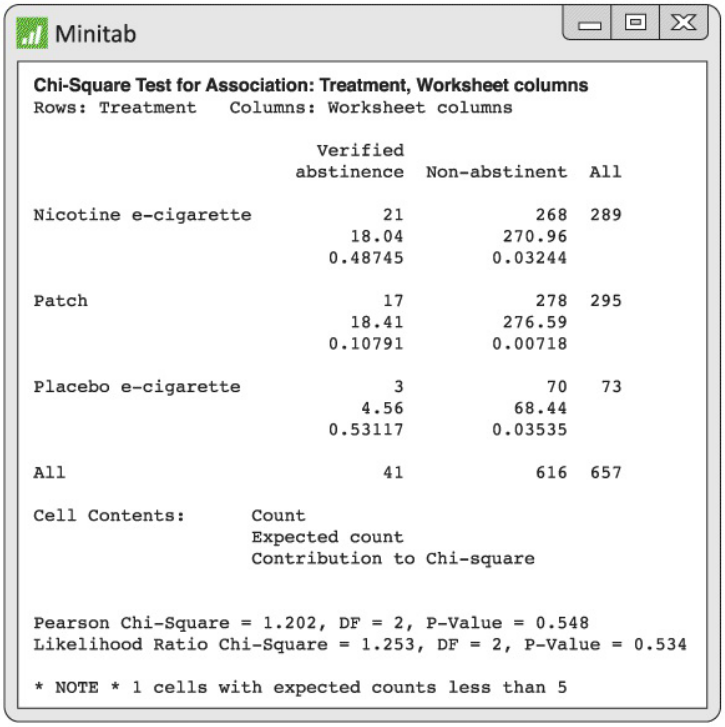
FIGURE 22.2 Output from the TI-83 graphing calculator, Minitab, JMP, and R Commander

for the chi-square test in the thromboembolism study.

APPLY YOUR KNOWLEDGE

22.5 E-cigarettes and smoking cessation, continued. Exercise 22.1 describes an experiment comparing three approaches for quitting smoking. Figure 22.3 shows the Minitab output for a chi-square test with a null hypothesis of no association between treatment received and abstinence outcome in the population of smokers who want to quit.

a. Show how to obtain the expected counts of smokers with or without verified abstinence at six months among those using a nicotine e-cigarette. Your results should agree, up to rounding error, with Figure 22.3. Verify that these two expected counts add to the row total for the observed counts in the nicotine e-cigarette group.
b. Compare the observed and expected counts in Figure 22.3. What do they suggest?
c. Show how the first chi-square component in the Minitab output (0.48745) was computed.
d. What is the value of the chi-square statistic reported by Minitab? Verify that, up to rounding error, this value is the sum of all six chi-square components shown in the Minitab output.



Chi-Square Test for Association: Treatment, Worksheet columns

```
Rows: Treatment   Columns: Worksheet columns

                                Verified
                              abstinence  Non-abstinent  All

Nicotine e-cigarette                  21            268  289
                                   18.04         270.96
                                 0.48745        0.03244

Patch                                 17            278  295
                                   18.41         276.59
                                 0.10791        0.00718

Placebo e-cigarette                    3             70   73
                                    4.56          68.44
                                 0.53117        0.03535

All                                   41            616  657

Cell Contents:      Count
                    Expected count
                    Contribution to Chi-square


Pearson Chi-Square = 1.202, DF = 2, P-Value = 0.548
Likelihood Ratio Chi-Square = 1.253, DF = 2, P-Value = 0.534

* NOTE * 1 cells with expected counts less than 5
```

Figure 22.3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.3 Minitab chi-square output for the two-way table of e-cigarettes and smoking cessation.

22.6 Aphid righting behavior, continued. Exercise 22.2 describes a study of aphids dropping from various heights to escape predators. The null hypothesis of "no relationship" says that in the population of aphids, the proportions that land upright are the same when the dropping height is 3, 5, 10, and 20 cm.

a. Find the expected cell counts if this hypothesis is true and display them in a two-way table. Add the row and column totals to your table and check that they agree with the totals for the observed counts. Explain why some of the expected counts have the same value.

b. Are there any large deviations between the observed counts and the expected counts? What kind of relationship between the two variables do these deviations point to?

THE CHI-SQUARE TEST

Hypothesis tests require a set of null and alternative hypotheses, a test statistic, and a *P*-value that gives the probability if the null hypothesis was true of obtaining a test statistic at least as extreme as that computed. To obtain the *P*-value for a chi-square test with a null hypothesis of no relationship between two categorical variables, we need to have a known probability distribution for the test statistic, X2.χ^2. Deriving this distribution mathematically is beyond the scope of an introductory statistics textbook, so we will simply state the facts. When the null hypothesis is true, the X2χ^2 statistic follows a **chi-square distribution.**

In Chapter 21 we saw that the chi-square distributions are a family of distributions that take only positive values and are skewed to the right. A specific chi-square distribution is specified by giving its *degrees of freedom*. Figure 21.1 (page 530) shows three examples of chi-square distributions. Because chi-square distributions take only positive values, the more extreme values are always found on the right tail of the curve.

P-VALUE OF THE CHI-SQUARE TEST FOR TWO-WAY TABLES

In a chi-square test for a two-way table with rr **rows** and cc **columns,** when the null hypothesis is true, the test statistic X2 χ^2 has the chi-square distribution with (r-1)(c-1)$(r-1)(c-1)$ **degrees of freedom.**

The *P*-value of the chi-square test is the area to the right of the calculated X2χ^2 statistic under this chi-square distribution.

Software usually finds *P*-values for us. Table D in the back of the book gives critical values for several chi-square distributions and can be used to find approximate *P*-values when no software is available.

Figure 22.2 shows that the *P*-value for the chi-square test for the thromboembolism data is very close to zero, approximately 0.00009. (Notice that the three statistical programs refer to this test as **Pearson chi-square.**) Because this two-way table has 4 rows (r=4)$(r=4)$ and 2 columns

$(c = 2)$, the *P*-value is the area to the right of $X^2 = 21.268$ under the chi-square distribution with df = $(r-1)(c-1) = (3)(1) = 3$. The TI-83 calculator function X2cdf(21.268,1E99,3) and the =CHISQ.DIST.RT(21.268,3) command in Excel confirm that the *P*-value is indeed 9.3E−5 (or 0.000093). To use Table D instead, you need to compare $X^2 = 21.268$ to the critical values in the df = 3 row of Table D. We see that $X^2 > 17.73$, which indicates that the test *P*-value must be smaller than 0.0005.

df = 3

p	.0005
X^*	17.73

With a *P*-value so small, we reject the null hypothesis and conclude that the study provides very strong evidence that there is *some* relationship between the distributions of ERT history and a first diagnosis of thromboembolism in the population of postmenopausal women.

The chi-square test is the overall test for detecting relationships between two categorical variables. **When the test is statistically significant,** it is important to look at the data to assess the nature of the relationship. We have three ways to look at the thromboembolism data from a descriptive standpoint:

- **Compare appropriate percents:** Which outcomes occur in quite different percents of case- and control-subjects? This is the method we learned in Chapter 5.
- **Compare observed and expected cell counts:** Which cells have more or fewer observations than we would expect if H_0 was true?
- **Look at the components of the chi-square statistic:** Which cells contribute the most to the value of X^2?

EXAMPLE 22.5 ERT and thromboembolism: conclusions

The study found very strong evidence of a relationship between ERT history and a first diagnosis of thromboembolism in the corresponding populations of postmenopausal women $(P = 0.00009)$. That is, the distributions of ERT history are significantly different in the populations of postmenopausal women with or without a first diagnosis of thromboembolism.

All three ways of comparing the distributions show that the main difference is that a higher proportion of women diagnosed with a thromboembolism currently take ERT orally. This is noticeable in Figure 22.1, where the bars for “current oral ERT use” are most different for the cases and the controls. We also see that more than half the value of X2X^2 (13.0795 out of 21.268) comes from just one cell—the cell for thromboembolism cases and current oral ERT use. In this cell, the observed count is 32, which is nearly twice as many women as we would have expected (17.06) to find if the two variables had been independent.

The broader conclusion, however, is more challenging. The study is observational. That is, the association revealed in the analysis does not show that oral ERT use “causes” a higher rate of thromboembolism. Perhaps lurking variables are responsible for both a higher rate of oral ERT prescription and a higher rate of thromboembolism. In addition, the study has a retrospective, case-control design. As we discussed in Chapter 6, a major challenge of such studies is to find cases and controls that are as similar as possible in every respect except for the variable studied. The researchers did take great care in their sample selection. The 381 controls were enrolled at the same medical centers and at the same time as the cases, and they were selected to be similar in age (mean 62.1 years for the cases and 62.0 for the controls) and overall health (body mass index, hypertension, diabetes). A thorough analysis and discussion of possible confounding variables were also included in the research paper.

There is an important message here: The comprehensive examination of the potential weaknesses of a study is often the most challenging aspect of statistics—and maybe the most critical.

STATS
IN YOUR
WORLD

HE STARTED IT!

A study of deaths in bar fights showed that in 90% of the cases, the person who died started the fight. You shouldn’t believe this. If you killed someone in a fight, what would you say when the police ask you who started the fight? After all, dead men tell no tales.

APPLY YOUR KNOWLEDGE

22.7 E-cigarettes and smoking cessation, continued. Figure 22.3 shows the Minitab output for a chi-square test with a null hypothesis of no association between treatment received and abstinence outcome in the population of smokers who want to quit for the study first described in Exercise 22.1.

a. What is the *P*-value of the chi-square test?

b. Explain why it doesn’t make sense in this study to find the components that contribute the most to X2χ^2 or to examine closely how the observed counts differ from the expected counts.

c. Conclude in the context of this problem.

22.8 Aphid righting behavior, continued. In Exercises 22.2 and 22.6 you began to analyze data on the landing posture of aphids dropping from various heights to escape predators.

a. Starting from the observed and expected counts, find the eight components of the chi-square statistic and then the statistic X2χ^2 itself.

b. Software gives X2=11.2554$\chi^2 = 11.2554$ with a *P*-value of 0.01204 for the test. Which cells contribute the most to X2?χ^2? What kind of relationship do these components in combination with the column percents in the table suggest?

c. Conclude in the context of this problem.

CONDITIONS FOR THE CHI-SQUARE TEST

The chi-square test, like the z procedures for comparing two proportions, is an approximate method that becomes more accurate as the counts in the cells of the table get larger. We must therefore check that the counts are large enough to trust the *P*-value. Fortunately, the chi-square approximation is accurate for quite modest counts. Here is a practical guideline.[4]

CELL COUNTS REQUIRED FOR THE CHI-SQUARE TEST

You can safely use the chi-square test for a two-way table with r rows and c columns when *most of the* $r \times c$ **expected counts**—at least 80%, or 4 out of 5—have a value of 5.0 or greater and *all* $r \times c$ **expected counts** are 1.0 or greater. In particular, all four expected counts in a 2×2 table should be values equal to 5.0 or greater.

Note that the *P*-value of the chi-square test is the right-tail area under a chi-square distribution because the χ^2 test statistic has a chi-square distribution *when the null hypothesis is true.* This is why the requirement for the chi-square test involves *expected* counts rather than observed counts. The expected counts for the thromboembolism study of Example 22.1 appear in the Minitab output in Figure 22.2. The smallest expected count is 17.06, so the data easily meet the guideline for safe use of chi-square.

Statistical inference also requires that the data be representative of one or more target populations. Two-way tables can arise in several ways, however. The thromboembolism study compared the distribution of one categorical variable (ERT history) in two separate random samples in a case-control study. The next example illustrates a different situation involving just one random sample.

EXAMPLE 22.6 Dealing with online harassment

STATE: Online harassment, from name calling to threats, is unfortunately a fairly common experience for internet users. People who experience it either ignore it or respond (for example, by posting or by involving authorities). Is one method more effective than the other? A 2014 Pew Research Center survey of a random sample of American adults found that 455 of the 549 participants who had chosen to ignore online harassment felt that it was an effective way to deal with the issue. Of the 368 participants who had chosen to respond to online harassment, 276 felt that it had been effective.[5]

Fabiana Ponzi/Shutterstock.com

PLAN: The study involves two categorical variables, choice of reaction to harassment (ignore, respond) and perceived result (effective, not effective), recorded in a random sample of American adults who have experienced online harassment. The findings can be organized in a two-way table with 2 rows and 2 columns.

	Perceived result		
Reaction choice	**Effective**	**Not effective**	**Total**
Ignore	455	94	549
Respond	276	92	368

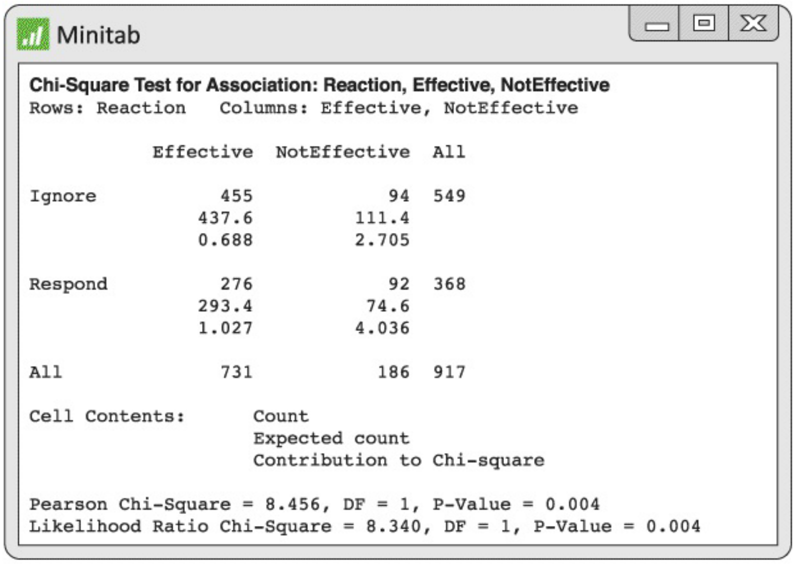
We will carry out a chi-square test for the following hypotheses:

H0:H_0 : There is no relationship between reaction choice and perceived result in the corresponding population (the two variables are independent).

Ha:H_a : There is some relationship between reaction choice and perceived result in the corresponding population (the two variables are not independent).

If the test is found to be statistically significant, we will describe the nature of the relationship by comparing row percents or observed versus expected cell counts or chi-square components.

SOLVE: First we check the guideline for use of chi-square. The expected cell counts appear in the Minitab output in Figure 22.4. The smallest expected count is 74.6, which definitely meets the requirement. The study uses a random sample. So we can safely use chi-square.



Chi-Square Test for Association: Reaction, Effective, NotEffective

```
Rows: Reaction   Columns: Effective, NotEffective

           Effective  NotEffective  All

Ignore           455            94  549
               437.6         111.4
               0.688         2.705

Respond          276            92  368
               293.4          74.6
               1.027         4.036

All              731           186  917

Cell Contents:      Count
                    Expected count
                    Contribution to Chi-square

Pearson Chi-Square = 8.456, DF = 1, P-Value = 0.004
Likelihood Ratio Chi-Square = 8.340, DF = 1, P-Value = 0.004
```

Figure 22.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.4 Minitab output for the two-way table in the online harassment study.

The output shows that the relationship is statistically significant (X2=8.456,$(X^2 = 8.456,$ P=0.004).$P = 0.004).$ The largest contributor to X2 X^2 comes from the cell for “respond” and “not effective” (4.0 out of 8.5). There are more individuals who felt that responding to harassment was not effective than we would expect if H0H_0 was true (92 observed but only 74.6 expected).

CONCLUDE: The survey found strong evidence (P=0.004)$(P = 0.004)$ that there is a relationship between choice of reaction and perceived result in the population of all American adults who experienced online harassment in the year 2014. That is, in this population, the perceived result (effective or not) depends significantly on the choice of reaction (ignore or respond). We find that 83% (455 out of 549) of the survey participants who chose to ignore the harassment felt that it had been an effective way to deal with it, compared with only 75% (276 out of 368) of those who had chosen to respond to it.

It is very important to acknowledge that this study is an observational survey and, therefore, we cannot conclude causation from it. Specifically, this study *does not* suggest that ignoring harassment is a more effective reaction than responding to it; any number of confounding variables could explain the observed association. For example, perhaps people who experience a milder form of harassment are more likely to ignore the harassment and perhaps milder harassment is more likely to resolve by itself. There is just no way to know the true nature of the observed relationship between reaction choice and result perception because the reaction was a personal “choice” influenced by a variety of factors (rather than a random assignment as part of a randomized experiment).

The computations performed by Minitab in this example involve the following steps:

- Expected counts, one for each cell in the table: expected count=row total × column totaltable total$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$
- Test statistic: X2=∑(observed count -expected count)2expected count $X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$
- Degrees of freedom of the chi-square distribution: (r-1)(c-1)=(2-1)(2-1)=1.$(r-1)(c-1) = (2-1)(2-1) = 1.$
- *P*-value as the area to the right of X2=8.456$X^2 = 8.456$ under the chi-square distribution with df=1.$df = 1.$ Software commands such as X2cdf(8.456,1E99,1) in the TI-83 calculator or =CHISQ.DIST.RT(8.456,1) in Excel give a *P*-value of 0.0036. Table D tells us only that the *P*-value must be between 0.0025 and 0.005 because the test statistic X2=8.456$X^2 = 8.456$ is between the critical values 7.88 and 9.14 listed in the df=1$df = 1$ row.

Notice one important way in which Examples 22.1 and 22.6 differ. In Example 22.6, unlike in Example 22.1, we do not have two separate random samples. Instead we have just one random sample of 917

American adults who have experienced online harassment, each one classified in two ways (by reaction choice and by result perception).

One of the most useful properties of chi-square is that it tests the null hypothesis of “no relationship between the row and column variables” whenever this hypothesis makes sense for a two-way table of counts. It makes sense when we are comparing a categorical response in two or more samples, as when we compared the history of ERT use for women with or without a diagnosis of thromboembolism. The hypothesis also makes sense when we have data on two categorical variables for the individuals in a single sample, as in Example 22.6. (In fact, the hypothesis of “no relationship” makes sense even if the single sample can be thought of as an entire population. Exercise 22.33 illustrates this case.)

USES OF THE CHI-SQUARE TEST

Use the chi-square test to test a null hypothesis of no relationship between two categorical variables, against a nondirectional alternative, when you have a two-way table from one of these situations:

- Independent SRSs from each of two or more populations, with each individual classified according to one categorical variable. (The other variable says which sample the individual comes from.)
- A single SRS, with each individual classified according to both of two categorical variables.

In fact, the chi-square test for two-way tables is sometimes called a **test of homogeneity** (“no difference” in distributions) when we compare the distribution of one categorical variable recorded in two or more random samples. It is also sometimes called a **test of independence** (the two variables are independent) when we examine the relationship between two categorical variables collected on the same random sample of individuals.

APPLY YOUR KNOWLEDGE

22.9 E-cigarettes and smoking cessation, continued. Figure 22.3 shows the Minitab output for a chi-square test with a null hypothesis of no association between treatment received and abstinence outcome in the population of smokers who want to quit for the study first described in Exercise 22.1.

a. Notice that Minitab provides the following warning: "* NOTE * 1 cells with expected counts less than 5." Explain why, despite this low expected count, the data in this study actually meet the cell count requirement for use of chi-square.
b. Did the study use a single SRS or multiple SRSs? What larger population or populations does the study aim to represent?

22.10 Aphid righting behavior, continued. Exercise 22.2 describes a study of aphids dropping from various heights to escape predators.

a. In Exercise 22.6 you computed the expected counts under a null hypothesis of no relationship between dropping height and landing side. Check whether the data meet the cell count requirement for use of chi-square.
b. Did the study use a single SRS or multiple SRSs? What larger population or populations does the study aim to represent?

22.11 CTE and contact sport. Chronic traumatic encephalopathy (CTE) is a progressive neurodegenerative disorder caused by repetitive brain trauma. A case-control study examined the brains of deceased men with a diagnosed non-genetic neurodegenerative disorder. The study compared individuals with and individuals without a history of participation in amateur contact sports (such as football, wrestling, or basketball). Of the 66 men who had participated in contact sport, 21 had CTE. None of the 132 control men who had not participated in a contact sport had CTE.[6]

a. Perform a chi-square test to see whether the study provides evidence that, among men with a neurodegenerative disorder, the percent with CTE is different for those with a history of contact sport participation and for those without such a history?

b. One of the observed counts is zero. Explain why the chi-square test is nevertheless appropriate for these data.

Rick Friedman/Sports Illustrated/
Getty Images

THE CHI-SQUARE TEST AND THE TWO-SAMPLE z TEST[i]

One use of the chi-square test is to compare the proportions of successes in any number of groups. When comparing the proportions of successes in only two groups, the data can be summarized in a 2×2 table like this one:

	Success	Failure
Group 1		
Group 2		

We then have two inference procedures available: the two-sample z test from Chapter 20 and the chi-square test with 1 degree of freedom for a 2×2 table. The null hypotheses tested are

$$H_0 : p_1 = p_2$$

for the two-sample z test, where p_1 and p_2 are the proportion of successes in each group, and

> H_0 : There is no relationship between the row and column variables.

for the chi-square test for two-way tables.

These two tests always agree. In fact, the chi-square statistic X^2 is just the square of the z statistic, and the *P*-value for X^2 is exactly the same as the two-sided *P*-value for z. We recommend using the z test to compare two proportions, because it gives you the choice of a one-sided test and is related to a confidence interval for the difference $p_1 - p_2$.

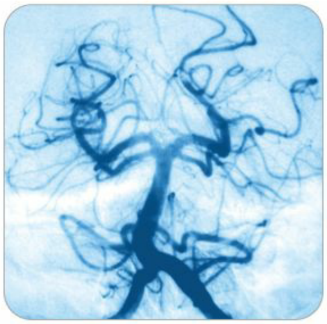
For instance, the data in Example 22.6 about online harassment can be organized into a 2×2 table. When we ran a chi-square test on these data, Minitab gave $X^2 = 8.456$ and $P = 0.004$. The 2-PropZTest function in the TI-83 calculator with a two-sided alternative reports $z = 2.908$ (positive or negative depending on which reaction choice is used for group 1) and $P = 0.0036$, which rounds to 0.004.

You can also verify that $X^2 = 8.456$ is the square value of $z = 2.908$.

APPLY YOUR KNOWLEDGE

22.12 Effectiveness of endovascular therapy. New surgical procedures are sometimes performed without evidence from randomized controlled experiments. Endovascular therapy has been increasingly used to supplement treatment with a clot-dissolving drug (t-PA) in patients with an acute ischemic stroke. It is not clear, however, if the addition of this surgical intervention truly benefits patients. A recent study randomly assigned patients with acute ischemic stroke to receive either a t-PA injection alone or a t-PA injection supplemented with endovascular therapy. Patients were examined 90 days later for evidence of functional independence. Of the 434 patients who received t-PA and endovascular therapy, 177 were functionally independent at 90 days, compared with 86 of the 222 patients who received t-PA alone.[7] We can test the hypothesis of “no difference” between the two groups in either of two ways: by using the two-sample z statistic or by using the chi-square statistic.

a. Check the conditions required for both tests, given in the boxes on pages 505 and 557. The conditions are very similar, as they ought to be.
b. State the null hypothesis with a two-sided alternative and carry out the z test. What is the test *P*-value?
c. Organize the data in a 2×2 table. Use the chi-square test to test the hypothesis from part b. Verify that the X^2 statistic is the square of the z statistic, and that the chi-square *P*-value agrees with the z result (up to the accuracy of the tables in the appendix if you do not use software).
d. What do you conclude about the effectiveness of supplementing t-PA with endovascular therapy in patients with an acute ischemic stroke?



Medical Body Scans/Science Source

CHAPTER 22 SUMMARY

- The **chi-square test** for a two-way table tests the null hypothesis H0H_0 that there is no relationship between the row variable and the column variable. The alternative hypothesis HaH_a says that there is some relationship but does not say what kind.

- The test compares the observed counts of observations in the cells of the table with the counts that would be expected if H0H_0 was true. The **expected count** in any cell is

  expected count=row total × column totaltable total

  $$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

  The **chi-square statistic** is

  X2=∑(observed count - expected count)2expected count

  $$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

- The *P*-value of the chi-square test for a two-way table of counts with rr rows and cc columns is the area to the right of the test statistic X2,X^2, under the **chi-square distribution** with (r-1)(c-1)$(r-1)(c-1)$ **degrees of freedom.** Use this *P*-value to decide whether to reject or fail to reject H0.H_0.

- The chi-square distribution is an approximation of the distribution of the statistic X2.X^2. You can safely use this approximation when all expected cell counts are at least equal to 1.0 and no more than 20% of them are less than 5.0.

- If the chi-square test finds a statistically significant relationship between the row and column variables in a two-way table, use descriptive data analysis to explain the nature of the relationship. You can do this by comparing well-chosen percents, comparing the observed counts with the expected counts, and looking for the largest **components of the chi-square statistic**.

THIS CHAPTER IN CONTEXT

In Chapter 19 we computed confidence intervals and tested hypotheses about the proportion p of successes in one population. In Chapter 20 we expanded the use of z procedures to situations where we want to compare the proportions of successes p_1 and p_2 in two distinct populations. In this chapter we describe how the chi-square procedure can be used to test the hypothesis that several populations have the same proportion of successes.

The chi-square procedures are very versatile and can also be used for categorical variables beyond the simple case of binary success/failure. In Chapter 21, we saw how the chi-square test for goodness of fit can be used to test the hypothesis that a population breaks down into set proportions of defined outcomes (the way a pie chart breaks down into its slices). In this chapter, we examine the use of the chi-square method for categorical data that can be represented in two-way tables. That is, we can break down the data according to two categorical variables, each with any number of levels (the r rows and c columns in the table).

We used two-way tables to describe relationships between two *categorical* variables in Chapter 5. We examined the marginal distribution of each variable and the conditional distribution of one variable among only those individuals who have a given value of the other variable. In Chapter 10 we addressed similar situations from a probability point of view and described the concept of independence. These ideas are directly relevant to the chi-square test for two-way tables, which is sometimes also called the chi-square test for independence.

As in Chapter 5, we must be careful not to assume that an observed association implies a cause-and-effect relation. A lurking variable may influence both variables studied and create the appearance of an association or change the nature of the association. Simpson’s paradox is an extreme example in which an association that holds for all of several groups reverses direction when these groups are combined into a single group.

CHECK YOUR SKILLS

Use the following information for Exercises 22.13 through 22.18. There are safe and effective interventions to treat respiratory failure in newborns, but they are typically costly and often not available in developing countries. Engineers designed a cheaper bubble continuous positive airway pressure (bCPAP) machine. In a low-resource area in Malawi, researchers compared the survival of infants with severe respiratory distress who received, at random, treatment with bCPAP or treatment with nasal oxygen therapy. They found that 44 of the 62 infants treated with bCPAP survived, compared with 11 of the 25 infants treated with nasal oxygen.[8] *We will test a null hypothesis of no relationship between survival outcome and treatment administered in the corresponding population.*

22.13 What percent of infants survived in the bCPAP treatment group?

a. 44%
b. 51%
c. 71%

22.14 The percent of infants surviving in the nasal oxygen treatment group is

a. higher than the percent of infants surviving in the bCPAP treatment group.
b. about the same as the percent of infants surviving in the bCPAP treatment group.
c. lower than the percent of infants surviving in the bCPAP treatment group.

22.15 What is, approximately, the expected count of surviving infants in the bCPAP treatment group?

a. 11
b. 39.2
c. 50.6

22.16 What is, approximately, the component of the chi-square statistic for the cell of surviving infants in the bCPAP treatment group?

a. 0.51
b. 0.59
c. 5.6

22.17 Software gives the chi-square statistic X2=5.572 $X^2 = 5.572$ for this two-way table. Using technology or Table D in the appendix, we can say that the *P*-value is

a. greater than 0.05.
b. between 0.01 and 0.05.
c. less than 0.01.

22.18 What can we reasonably conclude from the study findings?

a. Among infants with severe respiratory distress in a low-resource area of Malawi, the percent surviving is significantly different with the two treatments; infants treated with bCPAP have a higher survival rate.
b. The study failed to find significant evidence that the two treatments affect survival differently in the population of infants with severe respiratory distress in a low-resource area of Malawi.
c. The chi-square conditions for inference are not met, so a conclusion from this test is not reasonable.

Use the following information for Exercises 22.19 through 22.22. S-Adenosyl-L-methionine (SAMe) is a dietary supplement used to treat depression. A study randomly assigned smokers who wanted to quit smoking to one of three groups: a low dose of SAMe, a high dose of SAMe, or a placebo. After 8 weeks of treatment and behavioral smoking cessation intervention, smoking abstinence was assessed biochemically. Here are the findings:[9]

	Abstinent	**Still smoking**
Placebo	7	33
Low-dose SAMe	7	33
High-dose SAMe	5	35

22.19 What is the null hypothesis for the chi-square test for this two-way table?

a. There are equal proportions of subjects abstaining and still smoking in the placebo group.
b. There is no difference in the proportions of subjects abstaining from smoking among the three treatments.
c. There are equal numbers of subjects abstaining and still smoking.

22.20 What is the alternative hypothesis for the chi-square test for this two-way table?

a. The proportions of subjects abstaining from smoking are not the same for the three treatments.
b. The placebo group has the lowest proportion of subjects abstaining from smoking.
c. The high-dose SAMe group has the lowest proportion of subjects abstaining from smoking.

22.21 The most important fact that allows us to trust the results of the chi-square test is that

a. the study is large, 120 individuals in all.
b. no cell count is less than 5.
c. no expected cell count is less than 5.

22.22 Software gives X2=0.5003,P=0.7787$X^2 = 0.5003, P = 0.7787$ for this chi-square test. What can we reasonably conclude from the study findings?

a. Smokers who take the high dose of SAMe for 8 weeks are significantly more likely to abstain from smoking than smokers who take a lower dose or a placebo.
b. The study failed to find significant evidence that low-dose SAMe, high-dose SAMe, and a placebo lead to different proportions of subjects abstaining.
c. Smokers who take the high dose of SAMe for 8 weeks are significantly less likely to abstain from smoking than smokers who take a lower dose or a placebo.

CHAPTER 22 EXERCISES

22.23 Age and HPV infections. Human papillomavirus (HPV) infection is the most common sexually transmitted infection. Most people who become infected with HPV do not even know they have it. The body's immune system typically clears HPV naturally within two years, although serious health effects (such as cancer) can occur much later. The U.S. National Health and Nutrition Examination Survey (NHANES) contacted a representative sample of women between the ages of 14 and 59 years and asked them to provide a self-collected vaginal swab specimen. Here are the HPV test results broken down by age group:

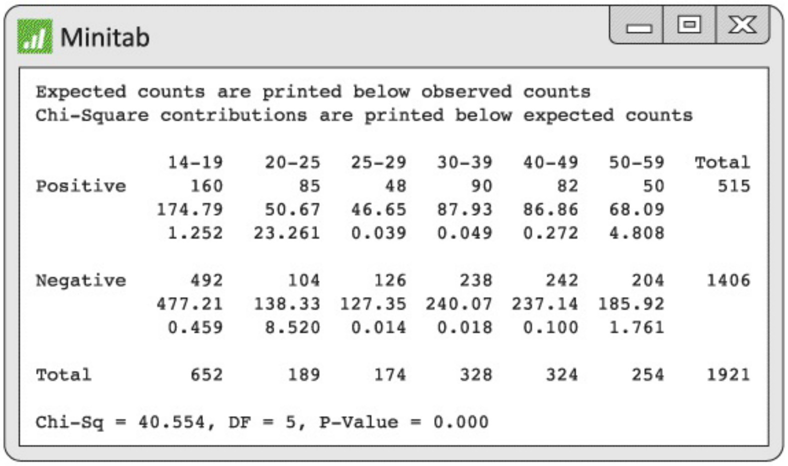
	Age group					
HPV test	**14–19**	**20–24**	**25–29**	**30–39**	**40–49**	**50–59**
Positive	160	85	48	90	82	50
Negative	492	104	126	238	242	204
Total	652	189	174	328	324	254

a. What percent of women who test positive for HPV are in the 14–19 age group? 20–24? 25–29? 30–39? 40–49? 50–59? Also find the conditional distributions of age, given a negative HPV test. Present the results in a table similar to the table of counts.

b. We want to see whether there is a relationship between HPV test and age group. The Minitab output for the corresponding chi-square test is displayed in Figure 22.5. Can we safely use the chi-square test?

c. According to Minitab, what are the chi-square statistic X2χ^2 and the *P*-value of the chi-square test? Does the study provide significant evidence that the proportion of women who test positive for HPV is not the same for all six age groups?

d. Look at the "Chi-square contributions" entries in Minitab's display. Which components contribute the most to X2?χ^2? Write a brief summary of the nature of the relationship between age group and HPV in women. Be sure to include some of the percents you computed in part a in your summary.



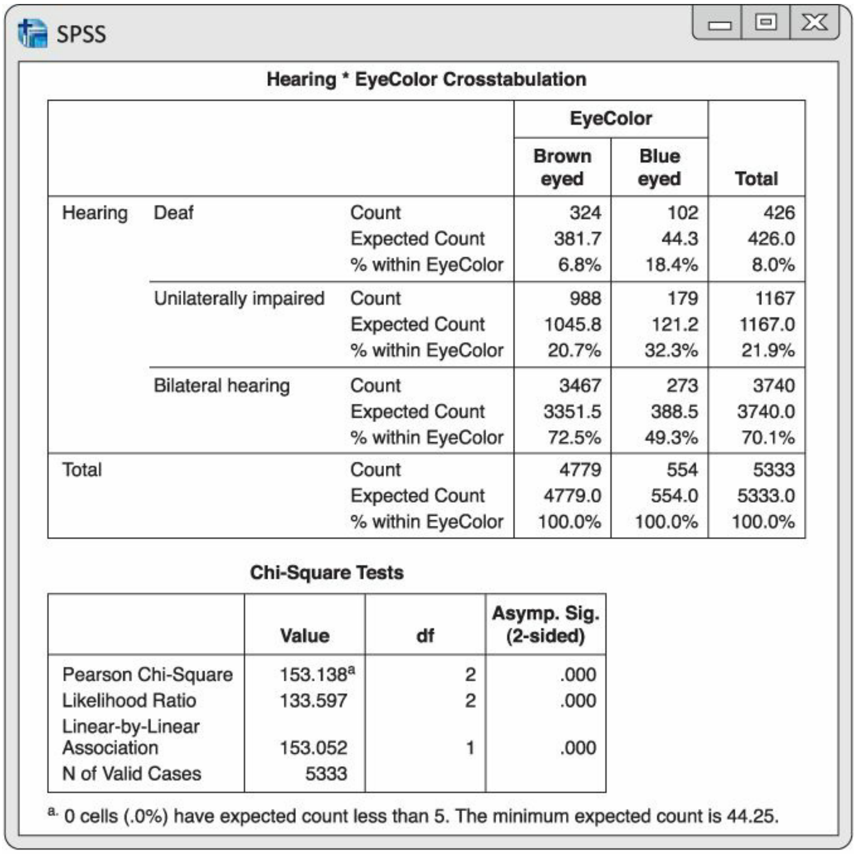
```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

            14-19   20-25   25-29   30-39   40-49   50-59   Total
Positive      160      85      48      90      82      50     515
           174.79   50.67   46.65   87.93   86.86   68.09
            1.252  23.261   0.039   0.049   0.272   4.808

Negative      492     104     126     238     242     204    1406
           477.21  138.33  127.35  240.07  237.14  185.92
            0.459   8.520   0.014   0.018   0.100   1.761

Total         652     189     174     328     324     254    1921

Chi-Sq = 40.554, DF = 5, P-Value = 0.000
```

Figure 22.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.5 Minitab output for the two-way table of HPV status by age group among American women.

22.24 Hearing impairment in Dalmatians. Congenital sensorineural deafness is the most common form of deafness in dogs and is often associated with congenital pigmentation deficiencies. A study of hearing impairment in dogs examined more than 5000 Dalmatians for both hearing impairment and iris color. Dogs with either one or both of their irises blue (a trait attributable to low iris pigmentation) were labeled "blue-eyed." Figure 22.6 shows the findings and the SPSS output for a chi-square test on these data.[10]

a. Describe the differences between the distributions of hearing impairments among brown-eyed and blue-eyed dogs with percents, with a bar graph, and in words.
b. Verify that the expected cell counts satisfy the requirement for use of chi-square.
c. Test the null hypothesis that there is no relationship between hearing impairment and iris color. Obtain the test *P*-value.
d. How do the observed and expected counts differ? Relate this to your conclusions in part a.



Hearing * EyeColor Crosstabulation

			EyeColor		
			Brown eyed	Blue eyed	Total
Hearing	Deaf	Count	324	102	426
		Expected Count	381.7	44.3	426.0
		% within EyeColor	6.8%	18.4%	8.0%
	Unilaterally impaired	Count	988	179	1167
		Expected Count	1045.8	121.2	1167.0
		% within EyeColor	20.7%	32.3%	21.9%
	Bilateral hearing	Count	3467	273	3740
		Expected Count	3351.5	388.5	3740.0
		% within EyeColor	72.5%	49.3%	70.1%
Total		Count	4779	554	5333
		Expected Count	4779.0	554.0	5333.0
		% within EyeColor	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	153.138[a]	2	.000
Likelihood Ratio	133.597	2	.000
Linear-by-Linear Association	153.052	1	.000
N of Valid Cases	5333		

[a] 0 cells (.0%) have expected count less than 5. The minimum expected count is 44.25.

Figure 22.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.6 SPSS output for the two-way table in the study of hearing impairment among Dalmatian dogs.

22.25 Peanut allergies in children. Peanut allergy is becoming increasingly common in Western countries. A study enrolled infants with a diagnosed peanut sensitivity and randomly assigned them to either avoid or regularly consume peanuts until 60 months of age. At the end of the study, 18 of the 51 children who had avoided peanuts were still allergic to peanuts. In contrast, 5 of the 47 children who had regularly consumed peanuts were still allergic to peanuts.[11] We want to run a test to see whether peanut allergy at 60 months is affected by the long-term consumption or avoidance of peanuts among children diagnosed in infancy with a peanut allergy.

a. Organize the study findings in a two-way table of counts. What are the proportions of children still allergic to peanuts at 60 months in the two feeding conditions?

b. Find the expected cell counts under the null hypothesis that there is no difference outcome between the two feeding conditions. Verify that the expected counts meet our guideline for use of the chi-square test.

c. Are the proportions of children still allergic to peanuts at 60 months significantly different for the two approaches? Obtain the test *P*-value and conclude in context.

22.26 Cranberry juice for preventing urinary tract infections. American cranberry (*Vaccinium macrocarpon*) was used medicinally by Native Americans for the treatment of bladder and kidney ailments. A randomized controlled study was designed to test whether regular drinking of cranberry juice can prevent the recurrence of urinary tract infections (UTIs) in women. In the study, 150 women with a urinary tract infection were treated with antibiotic and then randomly assigned to one of three groups. One group drank cranberry juice concentrate daily for six months; another group took a Lactobacillus drink daily for six months (Lactobacillus is a lactose-fermenting bacterium thought to help inhibit the growth of UTI-causing bacteria); the last group served as the control group and drank neither cranberry juice nor Lactobacillus drinks for six months. After six months, the number of women in each group with recurring symptomatic UTI (defined as one or more new infections) was recorded. Here are the results:[12]

	Outcome		
Treatment	**Recurring UTI**	**No new UTI**	**Total**
Cranberry juice	8	42	50
Lactobacillus drink	19	30	49
Control	18	32	50

a. Find the percent of women experiencing no new UTI in each of the three treatment groups. Make a graph to compare these percents. Describe the association between treatment and outcome.

b. Explain in words what the null hypothesis for the chi-square test says about treatment outcome. Verify that the conditions are met for a chi-square test.

c. Compute the chi-square statistic and find the *P*-value for this test. Examine the chi-square components to confirm the pattern you saw in part a. What is your overall conclusion? Interestingly, recent pharmacological studies have shown that some cranberry compounds prevent bacteria from binding to the membrane of host cells.

22.27 Stress and heart attacks. You read a newspaper article that describes a study of whether stress management can help reduce heart attacks. The 107 subjects all had reduced blood flow to the heart and so were at risk of a heart attack. They were assigned at random to one of three conditions: participation in a four-month stress management program, participation in a four-month exercising program, and a control condition consisting of the usual heart care from their personal physicians. The article goes on to say:

> *In the next three years, only three of the 33 people in the stress management group suffered "cardiac events," defined as a fatal or non-fatal heart attack or a surgical procedure such as a bypass or angioplasty. In the same period, seven of the 34 people in the exercise group and 12 out of the 40 patients in usual care suffered such events.*[13]

a. Use the information in the news article to make a two-way table that describes the study results. What are the success rates of the three treatments in preventing cardiac events?
b. Find the expected cell counts under the null hypothesis that there is no difference among the treatments. Verify that the expected counts meet our guideline for use of the chi-square test.
c. Is there a significant difference among the success rates for the three treatments? Obtain the test *P*-value and conclude in context.

22.28 HIV vaccine boost. Creating an effective HIV vaccine has been very challenging. A new approach combines a DNA vaccine with a delivery system using electrical impulses (electroporation) to boost the immune response. A preliminary study of the method's efficacy was conducted on adult volunteers. Each volunteer received one of five possible treatment options and a follow-up clinical examination to assess his or her immune response to the vaccine. Here is a two-way table that shows the results of the clinical exam after two immunization sessions:[14]

	Immune response	**No immune response**
0.2 mg vaccine via electroporation	3	5
1 mg vaccine via electroporation	7	1
4 mg vaccine via electroporation	6	2
4 mg vaccine intramuscularly	0	8
Saline solution via electroporation	0	8

We would like to know if the treatment type significantly influences the proportion of individuals showing an immune response. Can we use the chi-square test with these data? Check the conditions for inference and explain your answer.

22.29 Sex ratios and birth order in geese. Is sex determination truly a random process? The laws of genetics would suggest that it is because of the random distribution of X and Y chromosomes among sperm cells. However, physiological or environmental factors could favor one gender over the other under specific circumstances. A field biologist surveyed a random sample of lesser snow geese nests containing exactly four eggs. For each egg resulting in a live gosling, the laying order and the gender of the gosling were recorded. Here are the data:[15]

	Order in which egg is laid				
Sex	**1**	**2**	**3**	**4**	**Total**
Male	17	16	7	5	45
Female	10	9	17	14	50
Total	27	25	24	19	95

Is there evidence of a relationship between sex and laying order in lesser snow geese?

22.30 HPV and oropharyngeal cancer. Molecular data point to a possible link between the human papillomavirus (particularly

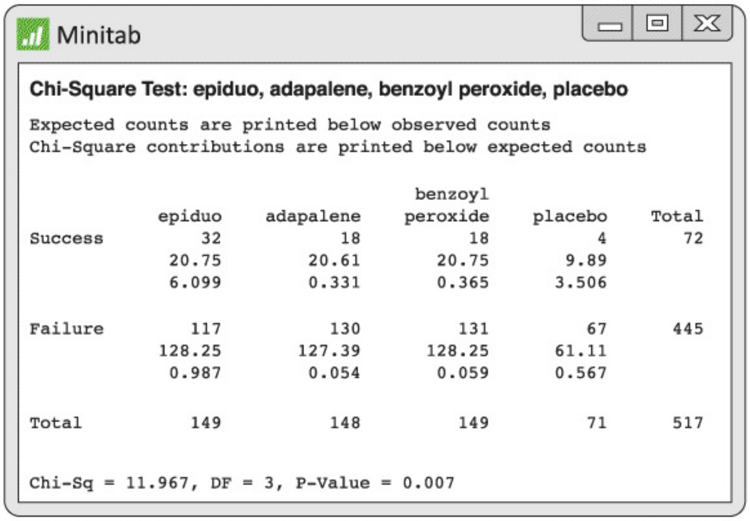
HPV-16) and oropharyngeal cancer. An epidemiologic study enrolled 100 hospitalized patients with newly diagnosed oropharyngeal cancer (cases) and 200 hospitalized patients without oropharyngeal cancer (controls) and determined whether they had been exposed to HPV-16. Here are the findings:[16]

	HPV-16 exposure	
	Yes	No
Oropharyngeal cancer	57	43
No oropharyngeal cancer	14	183

Is there significant evidence of a relationship between exposure to HPV-16 and oropharyngeal cancer? Can we conclude that HPV-16 causes oropharyngeal cancer?

22.31 Tongue piercing and cracked teeth. Researchers compared tooth health and periodontal damage in a group of 46 young adult males wearing a tongue piercing and a control group of 46 young adult males without tongue piercing. They found that 38 individuals in the tongue-piercing group and 26 in the control group had enamelcracks.[17] Does the study provide significant evidence of an association between enamel cracks and tongue piercing? Can you conclude that there is a causal relationship between tongue piercing and enamel cracks?

22.32 Comparing acne treatments. Acne is a common skin disease that affects most adolescents and can continue into adulthood. A study compared the effectiveness of three acne treatments and a placebo, all in gel form, applied twice daily for twelve weeks. The study’s 517 teenage volunteers were randomly assigned to one of the four treatments. Success was assessed as clear or almost clear skin at the end of the twelve-week period. Figure 22.7 gives counts and chi-square output for this study.[18] Is there significant evidence that the four treatments perform differently? If so, how do they compare?



Chi-Square Test: epiduo, adapalene, benzoyl peroxide, placebo

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

                                    benzoyl
           epiduo    adapalene     peroxide     placebo    Total
Success        32           18           18           4       72
            20.75        20.61        20.75        9.89
            6.099        0.331        0.365       3.506

Failure       117          130          131          67      445
           128.25       127.39       128.25       61.11
            0.987        0.054        0.059       0.567

Total         149          148          149          71      517

Chi-Sq = 11.967, DF = 3, P-Value = 0.007
```

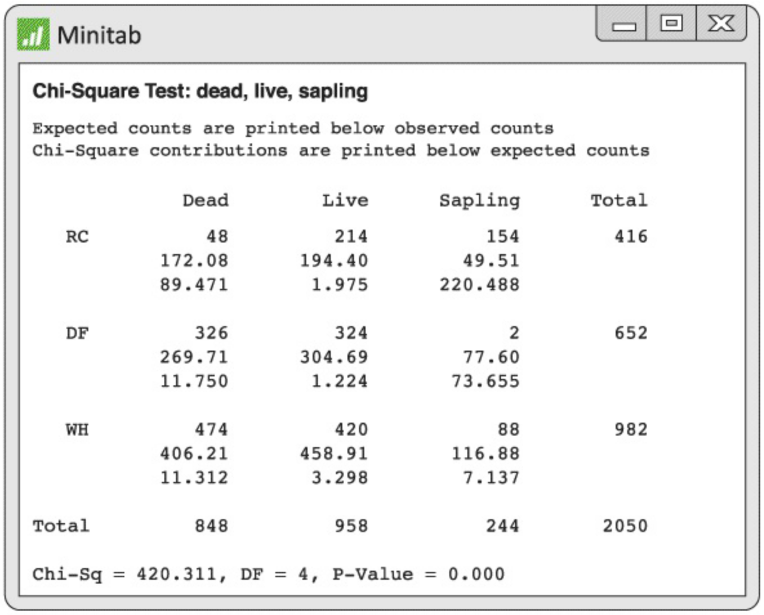
Figure 22.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.7 Minitab output for the study of acne treatments.

22.33 Characteristics of a forest. Forests are complex, evolving ecosystems. For instance, pioneer tree species can be displaced by successional species better adapted to the changing environment. Ecologists mapped a large Canadian forest plot dominated by Douglas fir with an understory of western hemlock and western red cedar. The following two-way table records all 2050 trees in the plot by species and by life stage. Sapling trees (young trees shorter than 1.3 meters) are indicative of the future of a forest.[19]

Tree species	Dead	Live	Sapling
Western red cedar (RC)	48	214	154
Douglas fir (DF)	326	324	2
Western hemlock (WH)	474	420	88

Do the data provide evidence that tree species and life stage are not independent in this forest area? Use the Minitab output in Figure 22.8. What does the study suggest about the state of this ecosystem?



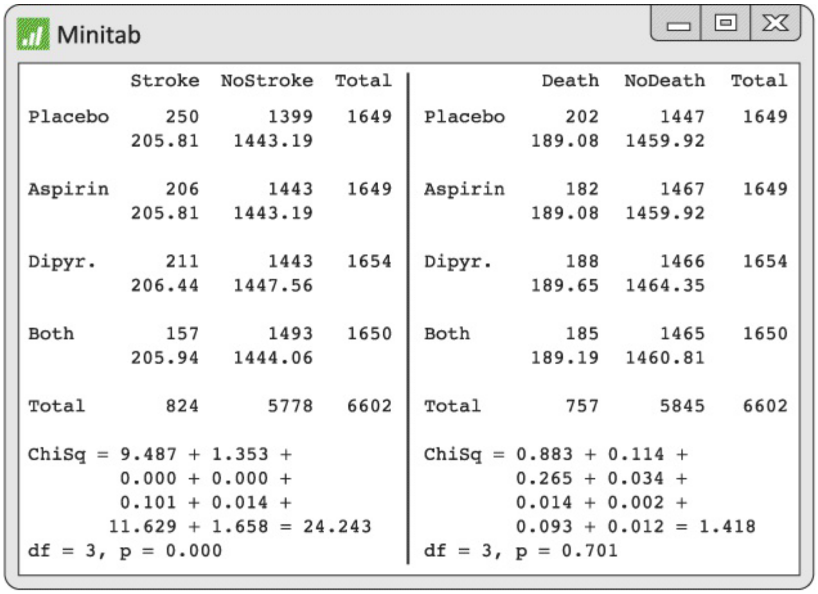
Chi-Square Test: dead, live, sapling

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

            Dead      Live    Sapling    Total

   RC         48       214        154      416
          172.08    194.40      49.51
          89.471     1.975    220.488

   DF        326       324          2      652
          269.71    304.69      77.60
          11.750     1.224     73.655

   WH        474       420         88      982
          406.21    458.91     116.88
          11.312     3.298      7.137

Total        848       958        244     2050

Chi-Sq = 420.311, DF = 4, P-Value = 0.000
```

Figure 22.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.8 Minitab output for the two-way table in the forest study.

22.34 Preventing strokes. Individuals who have experienced a stroke are at an increased risk of having another stroke. Prevention is particularly important for high-risk populations. The Second European Stroke Prevention Study asked whether daily doses of an anti-blood-clotting agent, such as aspirin or dipyridamole, would help prevent strokes among patients who had just experienced a first stroke. Volunteer patients were randomly assigned to one of 4 treatment options, including a placebo. Researchers recorded how many patients experienced another stroke and how many died of a stroke during the two years of the study. Figure 22.9 gives the observed and expected counts for this study, along with the chi-square calculations. Is there a relationship between treatment and strokes? Write a careful summary of your overall findings.



```
         Stroke  NoStroke  Total |          Death  NoDeath  Total
                                 |
Placebo     250      1399   1649 | Placebo    202     1447   1649
         205.81   1443.19        |         189.08  1459.92
                                 |
Aspirin     206      1443   1649 | Aspirin    182     1467   1649
         205.81   1443.19        |         189.08  1459.92
                                 |
Dipyr.      211      1443   1654 | Dipyr.     188     1466   1654
         206.44   1447.56        |         189.65  1464.35
                                 |
Both        157      1493   1650 | Both       185     1465   1650
         205.94   1444.06        |         189.19  1460.81
                                 |
Total       824      5778   6602 | Total      757     5845   6602
                                 |
ChiSq = 9.487 + 1.353 +          | ChiSq = 0.883 + 0.114 +
        0.000 + 0.000 +          |         0.265 + 0.034 +
        0.101 + 0.014 +          |         0.014 + 0.002 +
       11.629 + 1.658 = 24.243   |         0.093 + 0.012 = 1.418
df = 3, p = 0.000                | df = 3, p = 0.701
```

Figure 22.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.9 Minitab output for the experimental results of the study of stroke prevention.

22.35 More on preventing strokes. The study described in Exercise 22.34 also recorded the number of patients in each group who died of a stroke during the two years of the study. Is there a relationship between treatment and stroke-related deaths? Write a careful summary of your overall findings.

22.36 Do angry people have more heart disease? People who get angry easily tend to have more heart disease. That's the conclusion of a study that followed a random sample of 12,986 people from three locations for about four years. All subjects were free of heart disease at the beginning of the study. The subjects took the Spielberger Trait Anger Scale test, which measures how prone a person is to sudden anger. Here are data for the 8474 people in the sample who had normal blood pressure.[20] CHD stands for "coronary heart disease" (which includes both people who had heart attacks and people who needed medical treatment for heart disease).

	Low anger	**Moderate anger**	**High anger**	**Total**
CHD	53	110	27	190
No CHD	3057	4621	606	8284
Total	3110	4731	633	8474

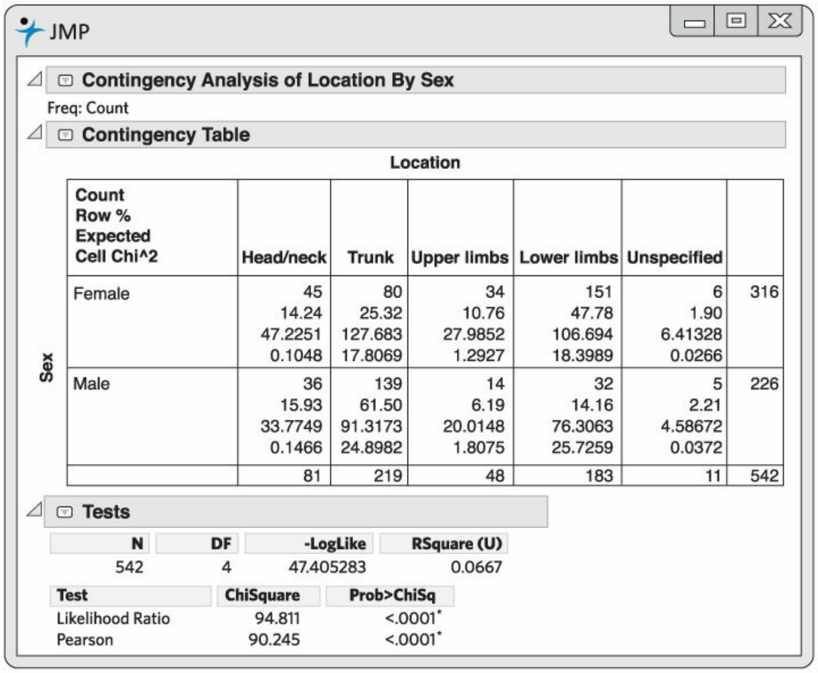
Do these data support the study’s conclusion about the relationship between anger and heart disease?

22.37 Echinacea for the common cold? On its website, the National Center for Complementary and Alternative Medicine (NCCAM, a branch of the NIH) offers comprehensive information about various complementary and alternative healing practices, including a summary of research findings for each practice. Echinacea is a flowering plant related to daisies and has a long history of medicinal use starting with Native Americans. Today it is the top-selling herbal remedy in the United States, mainly used against the common cold. In a double-blind experiment, healthy volunteers agreed to be exposed to common-cold-causing rhinovirus type 39. Three types of echinacea extracts were produced (labeled E1, E2, and E3 in the following table). Subjects received either an echinacea extract or a placebo (“P”) over two different periods: for seven consecutive days before viral exposure (infection and symptom prevention) and for seven consecutive days following exposure (symptom treatment). Among the variables studied were whether or not a subject became infected following viral exposure and whether or not infected subjects developed cold symptoms. Here are the results:[21]

Treatment	Infected	Not infected	Cold symptoms	No symptoms
E1 then E1	40	5	25	15
E2 then E2	42	10	24	18
E3 then E3	48	4	24	24
P then E1	43	5	27	16
P then E2	44	4	33	11
P then E3	44	7	28	16
P then P	88	15	58	30

Running a chi-square test for each variable separately gave X2=4.745 $X^2 = 4.745$ and X2=7.120$X^2 = 7.120$ for the first variable and the second variable, respectively. Write a careful summary of the overall findings from this study.

22.38 Melanoma in men and women. Intense or repetitive sun exposure can lead to melanoma, a rare but severe form of skin cancer. Patterns of sun exposure, however, differ between men and women. A study of cutaneous malignant melanoma in the Italian population recorded the distribution of melanoma by body location for men and for women.[22] Figure 22.10 displays the observed and expected counts for this study and the chi-square test from JMP. Write a careful summary of your overall findings.



```
Contingency Analysis of Location By Sex
Freq: Count
Contingency Table
                                   Location
     Count
     Row %
     Expected
     Cell Chi^2   Head/neck    Trunk  Upper limbs  Lower limbs  Unspecified
     Female              45       80           34          151            6   316
                      14.24    25.32        10.76        47.78         1.90
                    47.2251  127.683      27.9852      106.694      6.41328
                     0.1048  17.8069       1.2927      18.3989       0.0266
Sex  Male                36      139           14           32            5   226
                      15.93    61.50         6.19        14.16         2.21
                    33.7749  91.3173      20.0148      76.3063      4.58672
                     0.1466  24.8982       1.8075      25.7259       0.0372
                         81      219           48          183           11   542
Tests
       N    DF      -LogLike    RSquare (U)
     542     4     47.405283         0.0667
Test               ChiSquare   Prob>ChiSq
Likelihood Ratio      94.811      <.0001*
Pearson               90.245      <.0001*
```

Figure 22.10

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 22.10 JMP output for the melanoma epidemiological data.

22.39 Carcinogenicity of electromagnetic fields. The U.S. National Toxicology Program (NTP) studies the toxicity and carcinogenicity of agents potentially posing a risk to human health. Electromagnetic fields around electrical installations are harmless in theory, but they might have a long-term effect on health (for instance, in triggering tumors). Observational studies in human populations have suggested a potential association.

An NTP study looked at the occurrence of cancers in otherwise healthy rats after two years of daily exposure to 60-hertz electromagnetic fields of various intensities. Here are the numbers of male and female rats with a tumor after the two-year study period.[23] Note that 2 gauss is approximately 1000-fold greater than what is considered high exposure for humans.

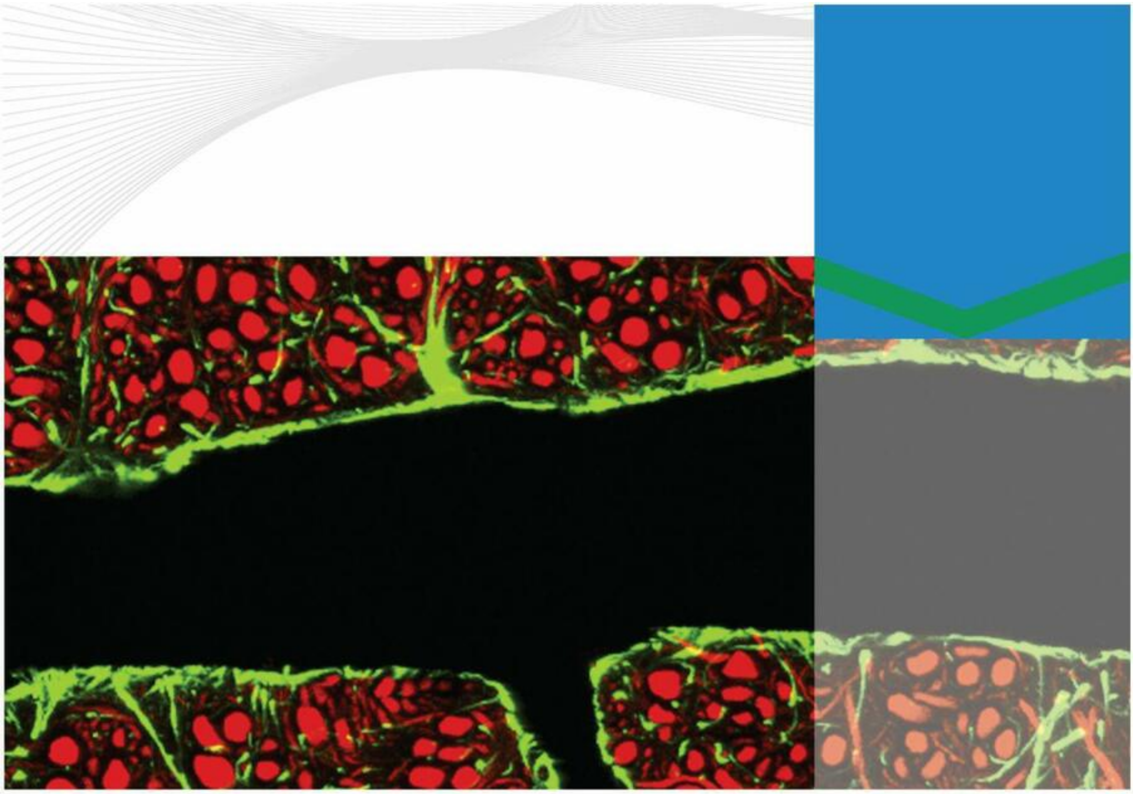
	Male rats	**Female rats**
Control: no exposure	16 (n = 99)	19 (n = 100)
0.02-gauss exposure	31 (n = 100)	22 (n = 100)
2-gauss exposure	30 (n = 100)	22 (n = 100)

In rats not exposed to an electromagnetic field (sample size in parentheses), is there significant evidence of a difference in the tumor rates between males and females?

22.40 Carcinogenicity study, continued. Use the data from Exercise 22.39 to answer the following questions.

a. Is there significant evidence of a relationship between tumor rate and electromagnetic exposure in male rats?
b. Is there significant evidence of a relationship between tumor rate and electromagnetic exposure in female rats?
c. Write a brief summary of your findings from both this exercise and Exercise 22.39.

CHAPTER 23 Inference for Regression



C.J. Guerin, PhD, MRC Toxicology Unit/Science Source

IN THIS CHAPTER WE COVER...

- The regression parameters
- Testing the hypothesis of no linear relationship
- Testing lack of correlation
- Confidence intervals for the regression slope
- Inference about prediction
- Checking the conditions for inference

When a scatterplot shows a linear relationship between a quantitative explanatory variable x and a quantitative response variable y, we can use the least-squares regression line fitted to the data to predict y for a given value of x.x. We discussed this approach in great detail in Chapters 3 and 4. When the data are a sample from a larger population, we need statistical inference to answer the following questions about the population:

- Is there really a linear relationship between xx and yy in the population, or might the pattern we see in the scatterplot plausibly arise just by chance?
- What is the slope (rate of change) that relates yy to xx in the population, including a margin of error for our estimate of the slope?
- If we use the least-squares line to predict yy for a given value of x,x, how accurate is our prediction (again, with a margin of error)?

This chapter shows you how to answer these questions. Here is an example we will explore throughout the chapter.

EXAMPLE 23.1 Aging and the blood–brain barrier

STATE: The blood–brain barrier (BBB) is a structure found around blood vessels serving the brain that limits which substances may enter the brain, thereby protecting the brain from toxins and pathogens. Researchers recruited 24 healthy adults of varying ages with no cognitive disability and measured BBB permeability in the hippocampus, a brain region involved in learning and memory. Because cognitive functions such as learning and memory typically decline with age, the study's objective was to examine the possible relationship between BBB permeability in the hippocampus and age. The findings are displayed in part of Table 23.1.[1] Age was measured in years and permeability was assessed with a high-resolution spatiotemporal imaging technique that yielded a "transfer constant" (Ktrans) in 10−3/min.

(K_{trans}) in $10^{-3}/\min$.

TABLE 23.1 Age (in years) and BBB permeability (in 10−3/min$10^{-3}/\min$) in t healthy adults. The table also includes computations (predictions residuals) from a least-squares regression analysis of the data.

Age	Permeability	Prediction	Residual y-	Residual^2 (y-	Age	Permeability	Predi

xx	yy	y^$\hat{y}$	y^$y-\hat{y}$	y^)2$(y-\hat{y})^2$	xx	yy	y
29	0.53	0.8770	− 0.3470	0.1204	65	1.85	1.2
28	0.81	0.8679	− 0.0579	0.0034	75	1.28	1.
30	0.83	0.8861	− 0.0561	0.0031	61	1.32	1.1
75	0.88	1.2950	− 0.4150	0.1722	80	1.36	1.3
76	0.98	1.3041	− 0.3241	0.1050	72	1.39	1.2
59	1.12	1.1496	− 0.0296	0.0009	84	1.40	1.3
55	1.09	1.1132	− 0.0232	0.0005	66	1.43	1.2
47	1.08	1.0405	0.0395	0.0016	91	1.43	1.4
23	1.02	0.8224	0.1976	0.0390	80	1.48	1.3
58	1.04	1.1405	− 0.1005	0.0101	90	1.51	1.4
40	1.05	0.9769	0.0731	0.0053	78	1.19	1.3
75	1.20	1.2950	− 0.0950	0.0090	78	1.22	1.3

PLAN: Make a scatterplot. If the relationship appears linear, use correlation and regression to describe it. Finally, ask whether there is a *statistically significant* relationship between age and BBB permeability.

SOLVE (FIRST PART): Chapters 3 and 4 introduced the data analysis that must come before inference. Let’s first review it. Figure 23.1 is a **scatterplot** of the data. Plot the explanatory variable (age) horizontally and the response variable (permeability) vertically. Look for the form, direction, and strength of the relationship, as well as for outliers or other deviations. There is a clear, though moderate, positive linear relationship, with possibly a mild outlier but no extreme outliers.

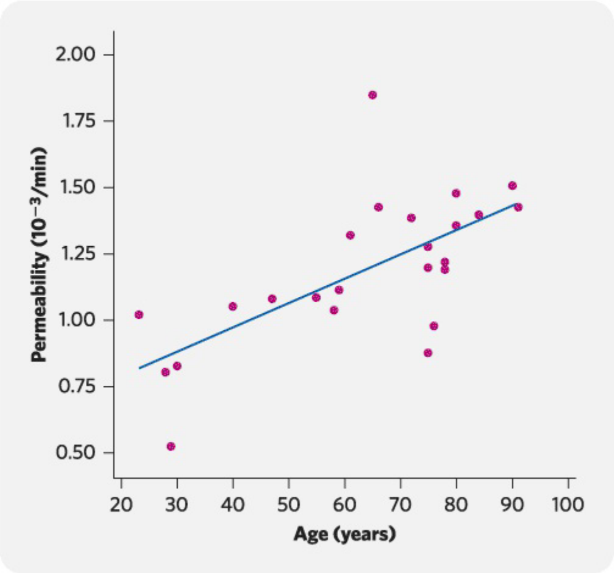


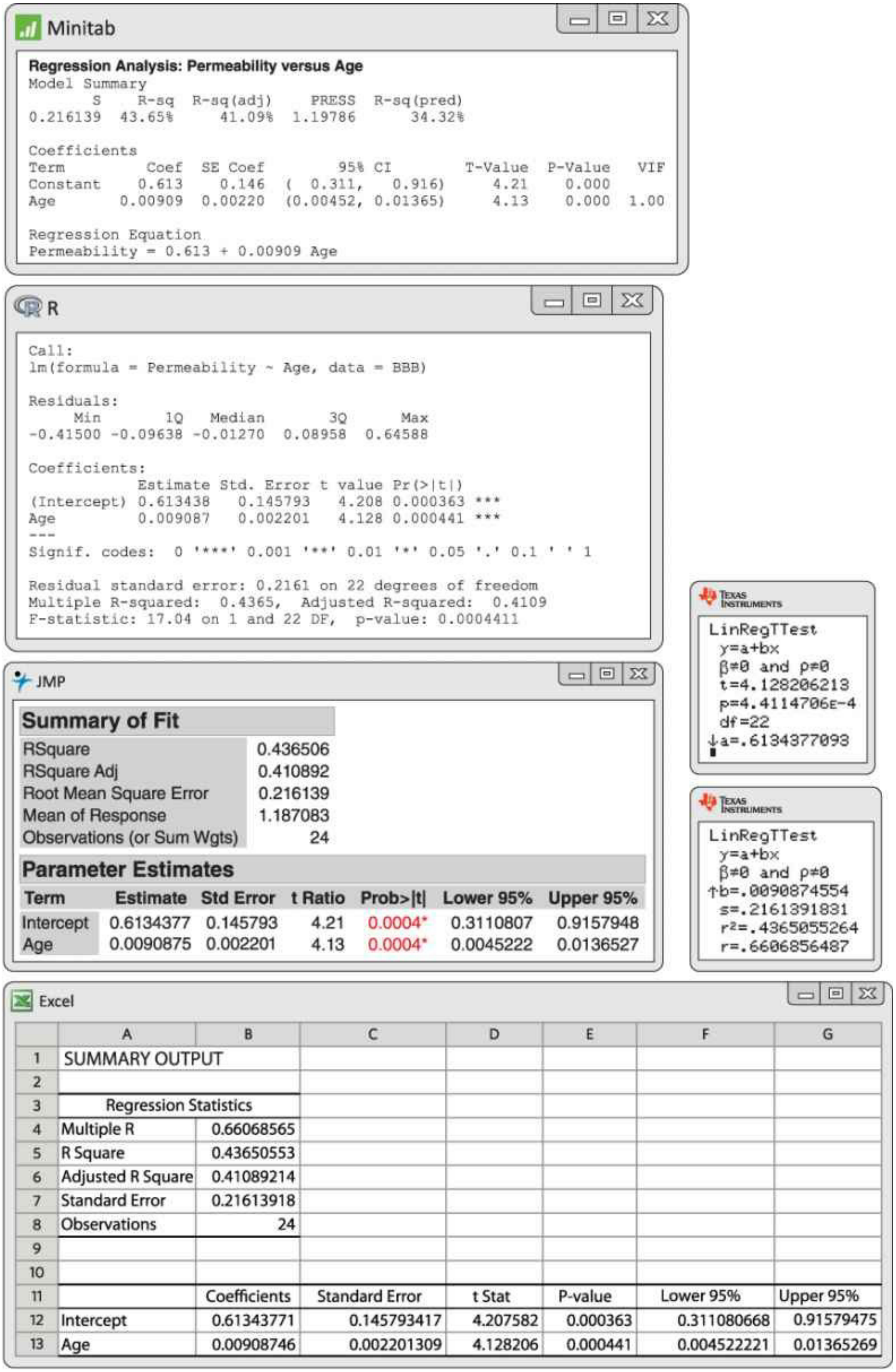
Figure 23.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 23.1 Scatterplot of blood-brain barrier (BBB) permeability in the hippocampus of 24 healthy adults as a function of their age.

Because the scatterplot shows a roughly linear (straight-line) pattern, the **correlation** describes the direction and strength of the relationship. The Excel output in Figure 23.2 indicates that the correlation between permeability and age is r=0.661.$r = 0.661$. We are interested in predicting the response from information about the explanatory variable, so we find the **least-squares regression line** for predicting permeability from age. When we entered the data into the statistical programs, we called the explanatory variable “Age”—a label used by most programs in Figure 23.2 to refer to the slope under the heading of “estimate” or “coefficient.” Thus the equation of the regression line is

y^=a+bx=0.6134+0.009087x

$$\begin{aligned}\hat{y} &= a + bx \\ &= 0.6134 + 0.009087x\end{aligned}$$



Regression Analysis: Permeability versus Age

```
Model Summary
       S    R-sq  R-sq(adj)    PRESS  R-sq(pred)
0.216139  43.65%     41.09%  1.19786      34.32%

Coefficients
Term          Coef  SE Coef          95% CI          T-Value  P-Value   VIF
Constant     0.613    0.146  (  0.311,    0.916)        4.21    0.000
Age        0.00909  0.00220  (0.00452, 0.01365)        4.13    0.000  1.00

Regression Equation
Permeability = 0.613 + 0.00909 Age
```

```
Call:
lm(formula = Permeability ~ Age, data = BBB)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41500 -0.09638 -0.01270  0.08958  0.64588

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.613438   0.145793   4.208 0.000363 ***
Age         0.009087   0.002201   4.128 0.000441 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2161 on 22 degrees of freedom
Multiple R-squared:  0.4365,  Adjusted R-squared:  0.4109
F-statistic: 17.04 on 1 and 22 DF,  p-value: 0.0004411
```

Summary of Fit

RSquare	0.436506
RSquare Adj	0.410892
Root Mean Square Error	0.216139
Mean of Response	1.187083
Observations (or Sum Wgts)	24

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	0.6134377	0.145793	4.21	0.0004*	0.3110807	0.9157948
Age	0.0090875	0.002201	4.13	0.0004*	0.0045222	0.0136527

```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
 t=4.128206213
 p=4.4114706E-4
 df=22
↓a=.6134377093
```

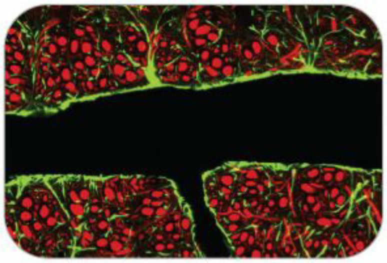
```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑b=.0090874554
 s=.2161391831
 r²=.4365055264
 r=.6606856487
```

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	Regression Statistics						
4	Multiple R	0.66068565					
5	R Square	0.43650553					
6	Adjusted R Square	0.41089214					
7	Standard Error	0.21613918					
8	Observations	24					
9							
10							
11		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
12	Intercept	0.61343771	0.145793417	4.207582	0.000363	0.311080668	0.91579475
13	Age	0.00908746	0.002201309	4.128206	0.000441	0.004522221	0.01365269

Figure 23.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 23.2 Regression of BBB permeability on age: output from a graphing calculator, three statistical programs, and a spreadsheet program.

CONCLUDE (FIRST PART): In this study, permeability of the BBB in the hippocampus increased linearly as a function of age. Because r2=0.437,$r^2 = 0.437$, approximately 44% of the variation in BBB permeability in the hippocampus is explained by the subject's age. We next ask whether this relationship, which is observed in a study of only 24 healthy adults, is statistically significant. We must now develop tools for inference in the regression setting.



C.J. Guerin, PhD, MRC Toxicology Unit/
Science Source

THE REGRESSION PARAMETERS

We can fit a least-squares regression line to *any* data relating two quantitative variables, though the results are useful *only* if the scatterplot shows a linear pattern. Statistical inference requires more detailed conditions. Because the conclusions of inference always concern some *population,* the conditions describe the population and how the data are produced from it. The slope b and intercept a of the least-squares line are *statistics*. That is, we calculated them from the sample data. These statistics would take somewhat different values if we repeated the analysis with different individuals. To do inference, think of a and b as estimates of unknown *parameters* that describe the entire population.

CONDITIONS FOR REGRESSION INFERENCE

We have n pairs of observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

- For any fixed value of x, the response y varies according to a **Normal distribution.** Repeated responses y are **independent** of each other.
- The mean response μ_y has a **straight-line relationship** with x given by a **population regression line**

$$\mu_y = \alpha + \beta x$$

  The slope β and intercept α are unknown parameters.

- The **standard deviation** of y (call it σ) is the same for all values of x. The value of σ is unknown.

Thus we must estimate three population parameters from the data: α, β, and σ.

Notice the distinction between x and y in the stated conditions. Unlike correlation, regression treats the two quantitative variables very differently. This distinction is particularly important for regression inference, because predictions are valid only for y,y, and the conditions for inference all concern the response variable y.y.

These conditions say that in the population there is “on average” a straight-line relationship between yy and x.x. The population regression line μy=α+βx$\mu_y = \alpha + \beta x$ says that the *mean* response μyμ_y moves along a straight line as the explanatory variable xx changes. We can’t observe the population regression line. The values of yy that we do observe vary according to a Normal distribution centered on their mean for each value of x.x. If we hold xx fixed and take many observations on y,y, the Normal pattern will eventually appear in a dotplot or histogram. In practice, we observe yy for many different values of x,x, so that we see an overall linear pattern formed by points scattered about the population regression line. The standard deviation σσ determines whether the points fall close to the population regression line (small σσ) or are widely scattered (large σσ).

Figure 23.3 shows the nature of regression data in picture form. The line in the figure is the population regression line. The mean of the response variable yy moves along this line as the explanatory variable xx takes different values. The Normal curves are the distribution of yy at different fixed values of x.x. All the curves have the same σ,σ, so the variability of yy is the same for all values of x.x. You should check the conditions for inference when you do inference about regression. We will see later how to do that.

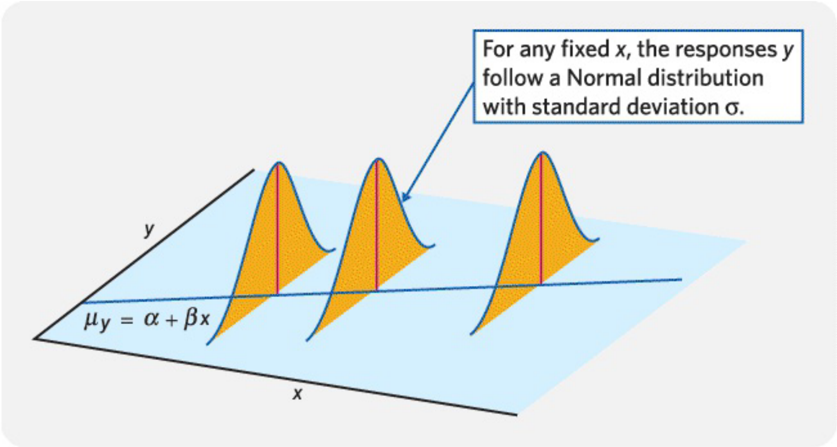
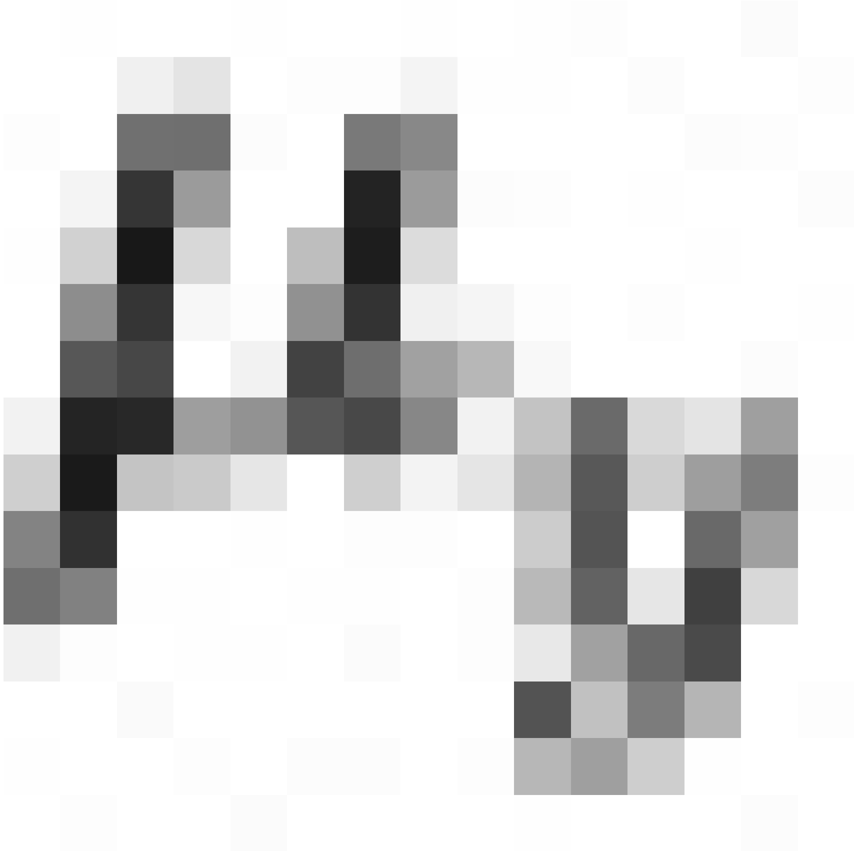

For any fixed x, the responses y follow a Normal distribution with standard deviation σ.
y
$\mu_y = \alpha + \beta x$
x

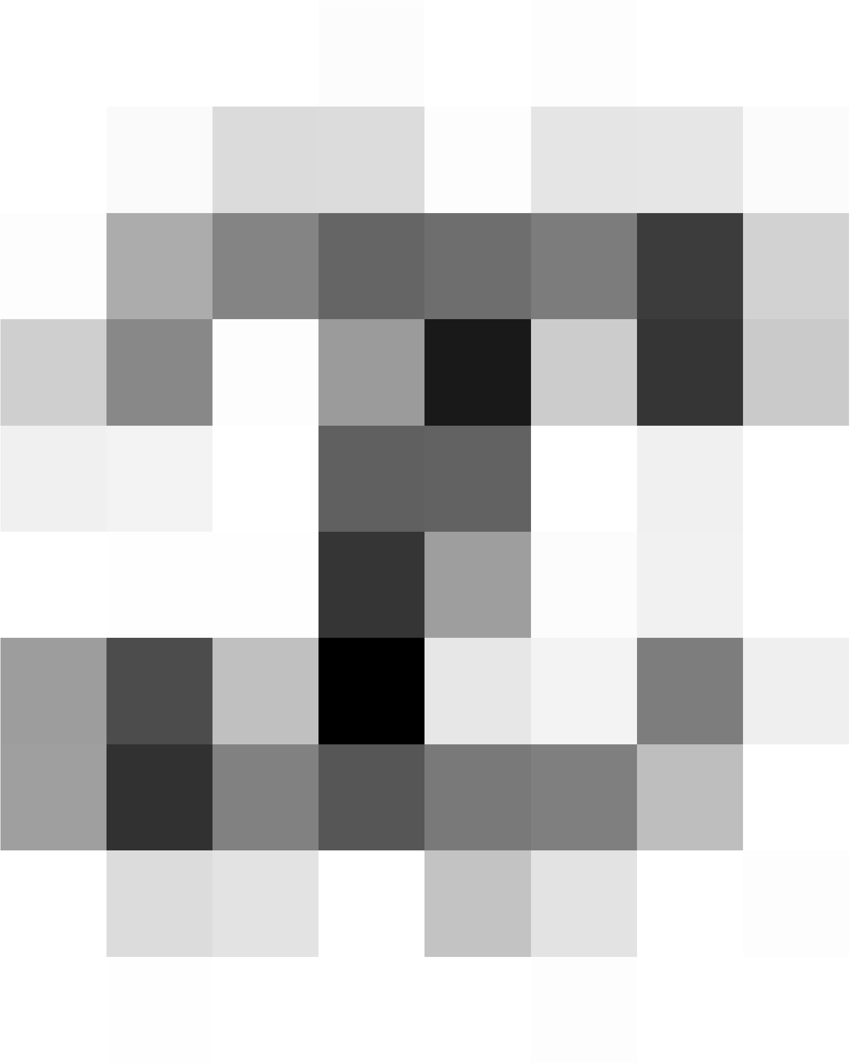
Figure 23.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 23.3 The nature of regression data when the conditions for inference are met. The line is the population regression line, which shows how the mean response μy

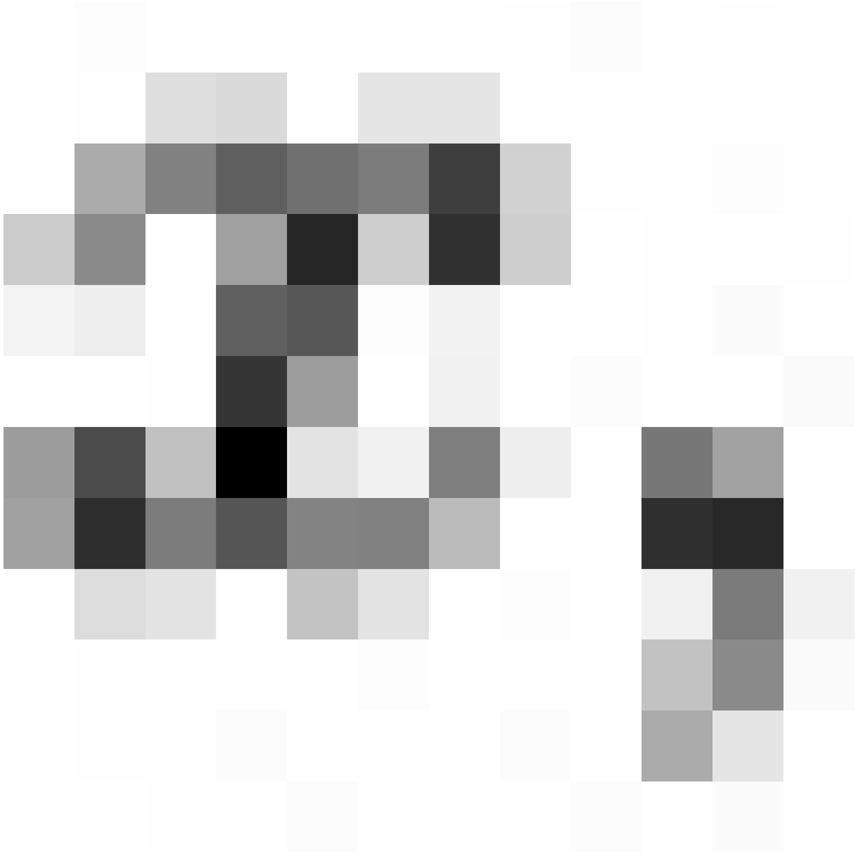


changes as the

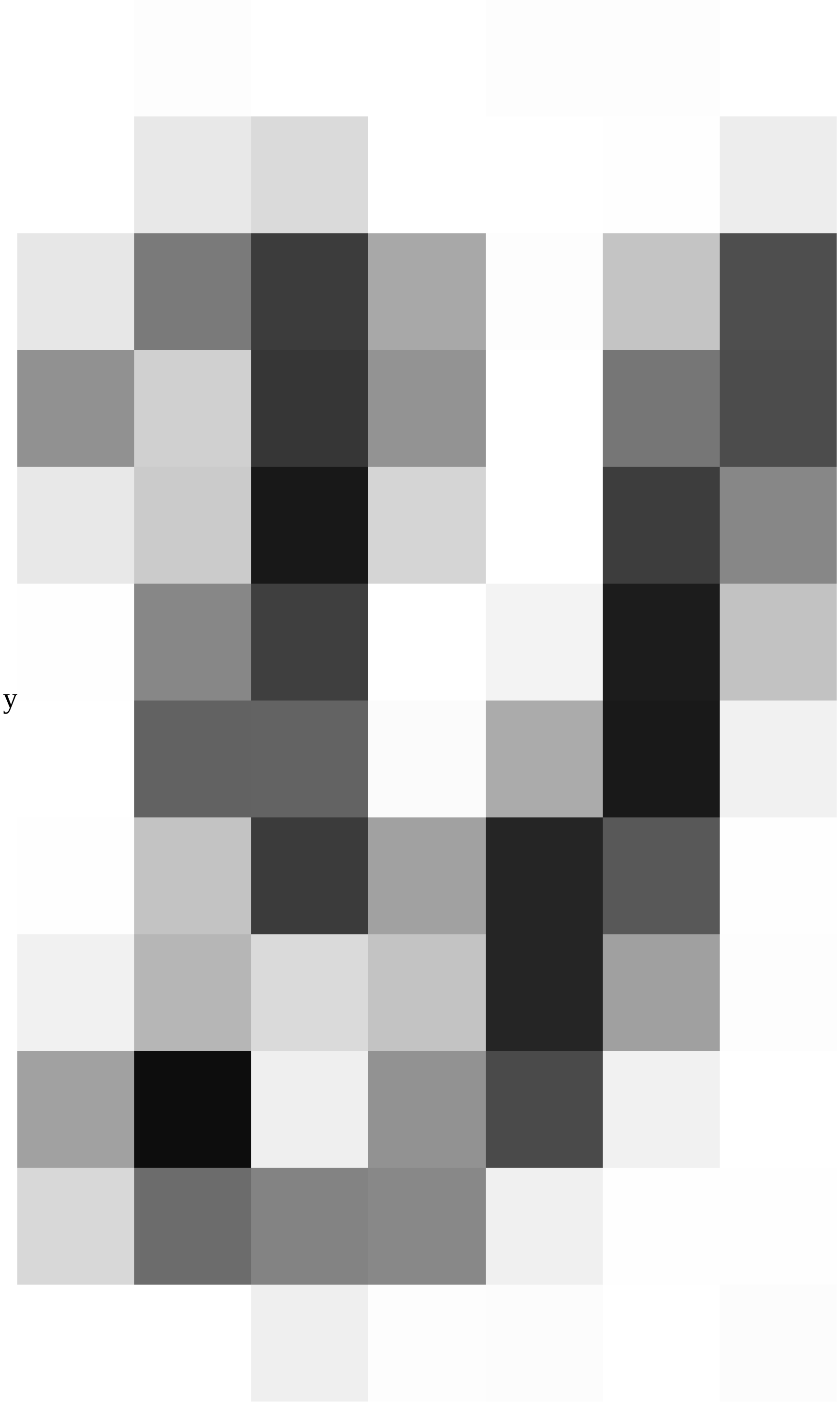
explanatory variable x



changes. For any fixed value of x,



the observed response

y  varies according to a Normal distribution having mean μy

μ and standard deviation σ.σ

When the conditions for inference are met, the first step is to estimate the unknown regression parameters α,α, β,β, and σ.σ. In Chapter 4, we described how the least-squares regression method provides an intercept a

a and a slope b.b. These values are estimates of the parameters αα and β,β, respectively.

The remaining parameter we need to estimate is the standard deviation σ,σ, which describes the variability of the response yy about the population regression line. Because the least-squares line estimates the population regression line, the **residuals** estimate how much yy varies about the population line. Recall that the residuals are the vertical deviations of the data points from the least-squares line:

residual=observed y-predicted y=y-y^

$$\begin{aligned}\text{residual} &= \text{observed } y - \text{predicted } y\\ &= y - \hat{y}\end{aligned}$$

There are nn residuals, one for each data point. Because σσ is the standard deviation of responses about the population regression line, we estimate it by a sample standard deviation of the residuals. We call this sample standard deviation the *regression standard error* to emphasize that it is estimated from data. We will study several kinds of inference in the regression setting. The regression standard error ss is the key measure of the variability of the responses in regression. It is part of the standard error of all the statistics we will use for inference in this chapter.

ESTIMATING THE REGRESSION PARAMETERS

When the conditions for regression are met and we calculate the least-squares line y^=a+bx,$\hat{y} = a + bx$, the **slope** bb of the least-squares line is an unbiased estimator of the population slope β,β, and the **intercept** aa of the least-squares line is an unbiased estimator of the population intercept α.α. The details of their computations were described in Chapter 4.

The **regression standard error** ss estimates the standard deviation σσ of responses about the mean represented by the population regression line. It is computed as follows:

s=1n-2∑residual2=1n-2∑(y-y^)2

$$s = \sqrt{\frac{1}{n-2}\sum \text{residual}^2} = \sqrt{\frac{1}{n-2}\sum (y-\hat{y})^2}$$

Any technology with basic statistical inference capabilities will perform all necessary computations and produce all three parameter estimates.

EXAMPLE 23.2 Aging and the blood–brain barrier: parameter estimates

The data in Figure 23.1, at first glance, satisfy the condition of scatter about an invisible population regression line. (We will carefully check these conditions later.) Using technology, we found that the least-squares line is y^=0.6134+0.009087x.$\hat{y} = 0.6134 + 0.009087x$. The slope is particularly important. *A slope is a rate of change.* The population slope ββ says how much more permeable, on average, the BBB in the hippocampus is with one additional year of age. Because b=0.009087 $b = 0.009087$ estimates the unknown β,β, we estimate that, on average, hippocampal BBB permeability is approximately 0.0090870×10-3/min $0.0090870 \times 10^{-3}/\text{min}$ higher for each additional year of age.

We need the intercept a=0.6134$a = 0.6134$ to draw the line, but it has no statistical meaning in this example. All study participants were adults, so we have no data near x=0.$x = 0$. This study's findings are not applicable to children and newborns, therefore the intercept here has a mathematical interpretation but no interpretation in context.

Lastly, we need the regression standard error ss to estimate the standard deviation σσ describing the variability of the response yy about the population regression line. All outputs in Figure 23.2 provide this information, though the labeling varies (JMP, in particular, uses the label "Root Mean Square Error," a term we will discuss in Chapter 24). Because s=0.2161$s = 0.2161$ estimates the unknown σ,σ, we estimate that the standard deviation of all residuals about the population regression line is approximately 0.2161×10-3/min.$0.2161 \times 10^{-3}/\text{min}$.

Wesaw in Chapter 4 how software computes the slope and the intercept of the least-squares regression line. The computations performed by software for the regression standard error ss involve the following steps, as detailed in Table 23.1:

- Residuals: y-y^,$y - \hat{y}$, computed for each of the 24 individuals in the data set. Note that the residuals from a least-squares line always have a mean of zero.
- Sum of squared residuals: ∑(y-y^)2=1.0277,$\sum(y-\hat{y})^2 = 1.0277$, as can be seen by adding the 24 values in the Residual^2 column of Table 23.1.
- Sample variance about the regression line: s2=∑(y-y^)2n-2=1.027722≈0.04671.$s^2 = \frac{\sum(y-\hat{y})^2}{n-2} = \frac{1.0277}{22} \approx 0.04671$. The value s2$s^2$ is an average of the squared deviations of the data points from the line, so it qualifies as a variance. We average the squared deviations by dividing by n-2,$n-2$, or 2 less than the number of data points. It turns out that if we know ss and n-2$n-2$ of the nn residuals, the other two residuals are determined. That is, n-2$n-2$ is the **degrees of freedom** of s2.s^2.
- Regression standard error: s=s2=0.04671≈0.2161.

$$s = \sqrt{s^2} = \sqrt{0.04671} \approx 0.2161.$$

Recall that we first encountered the idea of degrees of freedom in Chapter 2, when obtaining the standard deviation of a random sample of n n observations, which has n-1$n-1$ degrees of freedom. Now we observe two variables rather than one, and the degrees of freedom for regression are n-2$n-2$ rather than n-1.$n-1$.

APPLY YOUR KNOWLEDGE

23.1 Algal bloom control. Algal blooms can have negative effects on an ecosystem by dominating its phytoplankton communities. *Gonyostomum semen* is a nuisance alga infesting many parts of northern Europe. Could the overall biomass of *G. semen* be controlled by grazing zooplankton species? A research team examined the relationship between the net growth rate of *G. semen* and the number of *Daphnia magna* grazers introduced in test tubes. Net growth rate was computed by comparing the initial and final abundances of *G. semen* in the experiment, with a negative value indicative of a decrease in abundance. Here are the findings:[2]

Number of *D. magna* grazers	1	2	3	4	5	6

Net growth rate of *G. semen*	–1.9	–2.5	–2.2	–3.9	–4.1	–4.3

a. Examine the data. Make a scatterplot with number of grazers as the explanatory variable and find the correlation. There is a reasonably strong linear relationship.

b. Explain in words what the slope ββ of the population regression line would tell us if we knew it. Based on the data, what are the estimates of ββ and the intercept αα of the population regression line?

c. Calculate by hand the residuals for the six data points. Check that their sum is 0 (aside from roundoff error). Use the residuals to estimate the standard deviation σσ that measures variation in the responses (net growth rate) about their means given by the population regression line. You have now estimated all three parameters.

Justin Kase z12z/Alamy Stock Photo

23.2 Frog mating calls. *Hyla chrysoscelis* is a type of gray tree frog with a distinct mating call. Frogs are cold-blooded animals, and their physiology is affected by variations in temperature. Field biologists wanted to know if the mating call of *H. chrysoscelis* is affected by the temperature of its natural habitat. Here is the note-repetition frequency (in notes per second) of the mating call as a function of habitat temperature (in degrees Celsius, °C) for 20 *H. chrysoscelis* in the wild:[3]

Temperature (°C)	19	21	22	22	23	23	23	23	23	24

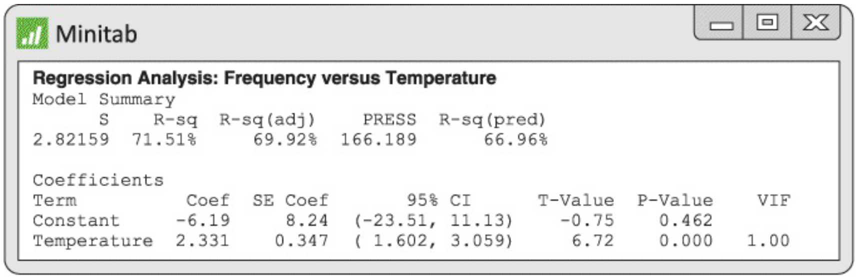
Frequency (notes/sec)	38	42	45	45	41	45	48	50	53	51
Temperature (°C)	24	24	24	25	25	25	25	26	26	27
Frequency (notes/sec)	48	53	47	53	49	56	53	55	55	54

We want to predict call frequency from habitat temperature. Figure 23.4 shows the Minitab regression output for these data.

a. Create a scatterplot suitable for predicting frequency from temperature. The pattern is linear. What is the squared correlation r2?r^2? Temperature explains a lot of the variation in call frequency.

b. For regression inference, we must estimate the three parameters α,α, β,β, and σ.σ. From the output, what are the estimates of these parameters?

c. What is the equation of the least-squares regression line of frequency on temperature? Add this line to your plot. We will continue the analysis of these data in later exercises.



Regression Analysis: Frequency versus Temperature

```
Model Summary
       S    R-sq  R-sq(adj)    PRESS  R-sq(pred)
2.82159  71.51%     69.92%  166.189      66.96%

Coefficients
Term         Coef  SE Coef       95% CI         T-Value  P-Value    VIF
Constant    -6.19     8.24  (-23.51, 11.13)     -0.75    0.462
Temperature  2.331    0.347  ( 1.602, 3.059)      6.72    0.000   1.00
```

Figure 23.4

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 23.4 Minitab output for the frog mating-call data.

Ryan M. Bolton/Alamy Stock Photo

TESTING THE HYPOTHESIS OF NO LINEAR RELATIONSHIP

Example 23.1 asked, “Is there a relationship between BBB permeability in the hippocampus and age?” That is, does permeability tend to change with age in this part of the brain? Data analysis supports this conjecture. But is the positive linear association found in this study of 24 healthy adults statistically significant? That is, is it too strong to have occurred just by chance? To answer this question, we test hypotheses about the slope ββ of the population regression line:

H0:β=0Ha:β≠0

$$\begin{aligned} H_0 &: \beta = 0 \\ H_a &: \beta \neq 0 \end{aligned}$$

A regression line with slope 0 is horizontal. That is, the mean of yy does not change at all when xx changes. So H0H_0 says that there is *no linear relationship* between xx and yy in the population. Put another way, H0H_0 says that *linear regression of* yy *on* xx *is of no value for predicting* yy. The alternative hypothesis for this study is two-sided because we do not know enough about the problem to expect that the relationship can be only in one particular direction (for example, perhaps permeability increases with age, which may result in increased toxins entering the brain, or perhaps permeability decreases with age, which may result in brain tissue that is starved of its essential nutrients).

The test statistic is just the standardized version of the least-squares slope b,b, using the hypothesized value β=0$\beta = 0$ for the mean of all sample bb values. It is simply another *t* statistic. Here are the details.

HYPOTHESIS TEST FOR REGRESSION SLOPE

Draw an SRS of nn pairs of observations on a quantitative explanatory variable xx and a quantitative response variable yy to ask whether they might have a linear relationship. To **test the hypothesis** H0:β =0,$H_0 : \beta = 0$, obtain the tt statistic

t=bSEb

$t = \frac{b}{\mathrm{SE}_b}$

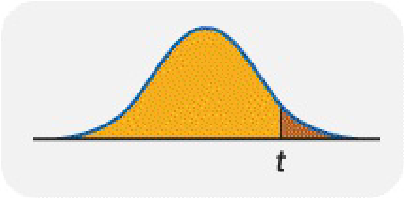
In this formula, the standard error of the least-squares slope bb is

SEb=s∑(x-x¯)2

$$\mathrm{SE}_b = \frac{s}{\sqrt{\sum (x - \overline{x})^2}}$$

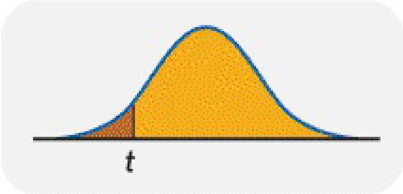
The sum runs over all observations on the explanatory variable x.x. In terms of a random variable TT having the t(n-2)$t\,(n-2)$ distribution, the *P*-value for a test of H0H_0 against

Ha: β>0$H_a: \beta > 0$ is P(T≥t)$P\,(T \geq t)$



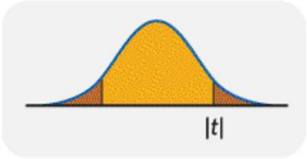
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Ha: β<0$H_a: \beta < 0$ is P(T≤t)$P\,(T \leq t)$



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Ha: β≠0$H_a: \beta \neq 0$ is 2P(T≥|t|)$2P\,(T \geq |t|)$



Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Any technology with basic statistical inference capabilities will perform all necessary computations (including SE_b) and provide both the test statistic and the *P*-value.

As always, the test *P*-value is the probability, if the null hypothesis was true, of obtaining a test statistic at least as extreme (in the direction of the alternative) as that computed. A small *P*-value provides enough evidence against H0H_0 to reject it in favor of Ha,H_a, leading us to call the results statistically significant, whereas a *P*-value that isn't small is inconclusive and we can only fail to reject H0.H_0.

EXAMPLE 23.3 Aging and the blood–brain barrier: Is the relationship significant?

The hypothesis H0: β=0$H_0 : \beta = 0$ says that age has no straight-line relationship with BBB permeability in the hippocampus. The alternative hypothesis is two-sided, simply stating that there is a linear relationship between xx and y,y, Ha: β≠0.$H_a : \beta \neq 0$.

All software outputs in Figure 23.2 give t=4.13$t = 4.13$ with a very small, highly statistically significant two-sided *P*-value of 0.0004 (which Minitab rounds to three decimal places as 0.000). There is very strong evidence that, on average, BBB permeability in the hippocampus increases with age in the population of healthy adults. Remember, however, that strong statistical significance does not imply a large effect or an important effect; rather, it simply indicates that the observed effect is highly unlikely to have arisen just by chance because of the random sampling process if the null hypothesis of no linear population relationship was true.

The computations performed by software in this example involve the following steps:

- Slope: b=0.00909.$b = 0.00909$.
- Standard error of the slope: SEb=0.00220.$SE_b = 0.00220$.
- Test statistic: $b/SE_b = 0.00909/0.00220 \approx 4.13$.

$b/\mathrm{SE}_b = 0.00909/0.00220 \approx 4.13.$

- Degrees of freedom of the *t* distribution: n-2=24-2=22.$n - 2 = 24 - 2 = 22.$
- *P*-value, computed as twice the area to the right of t=4.13$t = 4.13$ under the t(22)$t\,(22)$ distribution because the alternative is two-sided (so the *P*-value is the left-tail area plus the right-tail area). Software commands such as 2*pt (4.13, df=22, lower.tail=FALSE) in R or =T.DIST.2T(4.13, 22) in Excel give a *P*-value of 0.00044. Table C tells us only that the two-sided *P*-value must be less than 0.001 because the test statistic t=4.13$t = 4.13$ is greater than the largest critical value 3.792 listed for the t(22)$t\,(22)$ distribution.

df = 22

t*t^*	3.792
Two-sided *P*	.001

Most statistical programs give by default a two-sided *P*-value for regression inference, although some, like the TI-83 calculator, allow you to choose the direction of the alternative. When your test calls for a one-sided alternative, you can typically take the two-sided *P*-value provided and divide it by 2.

APPLY YOUR KNOWLEDGE

23.3 Algal bloom control: testing. Exercise 23.1 gives data on the number of *D. magna* grazers introduced in test tubes and the net growth rate of *G. semen* algae. Software tells us that the least-squares slope is b=-0.5286$b = -0.5286$ with standard error SEb=0.1059.$\mathrm{SE}_b = 0.1059.$

a. What is the *t* statistic for testing H0: β=0?$H_0 : \beta = 0?$
b. How many degrees of freedom does *t* have? Use software or Table C to obtain the test *P*-value against the one-sided alternative Ha: β<0$H_a : \beta < 0$ (as the grazers would be expected to reduce algal growth). What do you conclude?

Ted Kinsman/Science Source

23.4 Frog mating calls: testing. Exercise 23.2 gives data on the note-repetition frequency of the gray tree frog mating call as a function of habitat temperature, and Figure 23.4 shows the Minitab regression output for these data.

a. We want to know if call frequency varies linearly with habitat temperature for this species. Write the corresponding null and alternative hypotheses.
b. From the Minitab output, what are the test statistic and the *P*-value for this inference?
c. Conclude in context.

TESTING LACK OF CORRELATION[i]

In Chapter 4 we saw that the least-squares slope b is closely related to the correlation r between the explanatory and response variables x and y. In the same way, the slope β of the population regression line is closely related to the correlation between x and y in the population. In particular, the slope is 0 exactly when the correlation is 0.

Testing the null hypothesis $H_0: \beta = 0$ is therefore exactly the same as testing that there is *no correlation* between x and y in the population from which we drew our data. You can use the test for zero slope to test the hypothesis of zero correlation between any two quantitative variables. That's a useful trick.

Because correlation also makes sense when there is no explanatory-response distinction, it is handy to be able to test correlation without doing regression. The statistic r follows a particular distribution with $n - 2$ degrees of freedom. We don't cover the specifics of this distribution here, but Table E in the back of the book gives critical values of the sample correlation r under the null hypothesis that the correlation is 0 in the population. Use this table when *both variables* have at least approximately Normal distributions or when the sample size is large.

EXAMPLE 23.4 IQ and dyslexia in young children

STATE: IQ tests are sometimes used to assess children's cognitive abilities. However, typical IQ tests draw on reading skills, which may translate into lower scores for dyslexic children. Nonverbal IQ tests should circumvent this issue, allowing us to ask whether there is a relationship between reading skills and cognitive abilities in dyslexic children. Figure 23.5 displays the reading-skill scores (on a standardized scale) and the nonverbal IQ scores (also on a standardized scale) of 22 dyslexic children aged 7 to 10 years.[4] The relationship, if any, appears rather weak.

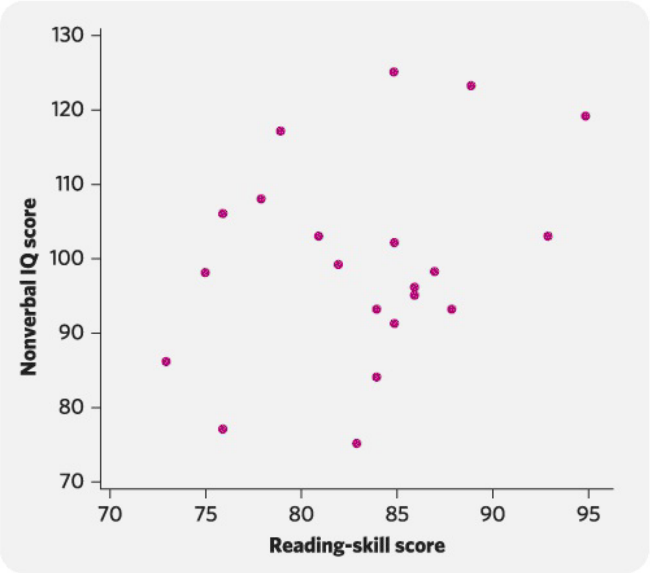


Figure 23.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 23.5 Scatterplot of the reading-skill scores and nonverbal IQ scores of 22 dyslexic children.

Historically, reading difficulties in children with low IQ test scores have often been attributed to the low IQ. However, recent research suggests that dyslexia has a neuronal foundation and tends to occur regardless of other cognitive abilities. Therefore, we want to test whether reading-skill score and nonverbal IQ score are significantly *correlated* in young dyslexic children.

PLAN: We test the hypotheses

H0: population correlation=0Ha: population correlation≠0

$$\begin{aligned} H_0 &: \text{population correlation} = 0 \\ H_a &: \text{population correlation} \neq 0 \end{aligned}$$

SOLVE: The data are displayed in the following table. Dotplots (we don't display them) show that the distributions of both variables are roughly symmetric, without outliers, and Normal quantile plots indicate that the deviations from Normality are minor. With 22 observations, a hypotheses test is safe.

Reading-skill score	84	81	85	82	87	85	89	86	85	83	76
Nonverbal IQ score	93	103	102	99	98	91	123	95	125	75	106

Reading-skill score	76	75	78	86	73	79	95	85	93	84	88
Nonverbal IQ score	77	98	108	96	86	117	119	91	103	84	93

The correlation coefficient for these data is $r=0.308$.$r = 0.308$. Compare this value with the critical values in the $n=20$$n = 20$ row of Table E. The closest critical values in that row are 0.2992 and 0.3783, for upper tail areas 0.10 and 0.05, respectively. HaH_a is two-sided, because we have no a priori expectation. Therefore, the *P*-value is $0.10<P<0.20$.$0.10 < P < 0.20$.

We could also have tested $H0:\beta=0$.$H_0 : \beta = 0$. The tt statistic for this test is $t=1.45$$t = 1.45$ with a two-sided *P*-value of 0.163 (given by software). **This is also exactly the *P*-value for** testing population correlation = 0.$\textbf{testing population correlation} = \mathbf{0}$.

CONCLUDE: The study failed to find evidence of a correlation between reading-skill scores and nonverbal IQ scores in all young dyslexic children. Absence of evidence is not evidence of absence, however, so we can never prove the null hypothesis. What the data do show is that, even if there was a relationship in the population, nonverbal IQ score would predict only about 9.5% of variations in reading-skill score, and vice versa (a rather weak predictor).

$n=20$$n = 20$

Critical value	0.2992	0.3783
Upper tail area	0.05	0.10
Two-sided *P*	0.10	0.20

APPLY YOUR KNOWLEDGE

23.5 Algal bloom control: testing correlation. Exercise 23.1 gives data on the number of *D. magna* grazers introduced in test tubes and the net growth rate of *G. semen* algae. There are only 6 observations, so we worry that the apparent relationship may be just chance. Is the correlation significantly less than 0? Answer this question in two ways.

a. Return to your t statistic from Exercise 23.3. What is the one-sided P-value for this t? Apply your result to test the correlation.

b. Find the correlation rr and use Table E to approximate the P-value of the one-sided test. Compare your answers using both methods.

23.6 Hand and body lengths. While the human body has overall proportions undeniably characteristic of the species, individuals do vary in size and body shape. Is there a relationship, for instance, between a person's height and hand size? Figure 23.6 is a scatterplot of the body lengths (in meters, m) and hand lengths (in millimeters, mm) of 21 healthy adult males, based on the following data.[5]

Body length (m)	**Hand length (mm)**	**Body length (m)**	**Hand length (mm)**	**Body length (m)**	**Hand length (mm)**
1.75	181	1.80	195	1.82	193
1.91	188	1.78	188	1.73	186
1.73	184	1.80	188	1.85	194
1.80	180	1.83	187	1.77	197
1.78	183	1.91	191	1.98	203
1.88	198	1.78	196	1.91	200
1.78	192	1.78	193	1.73	206

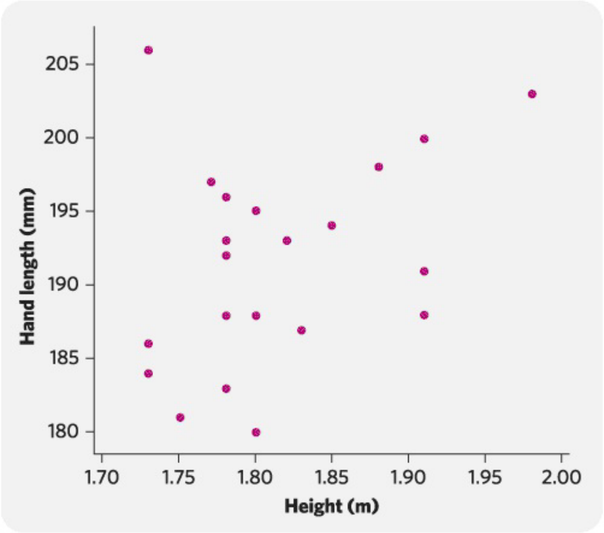


Figure 23.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 23.6 Scatterplot of hand lengths and body heights for 21 men.

Neither body length nor hand length can be considered a possible explanatory variable. Rather, genes and environmental factors influence both variables and are responsible for any relationship that may exist between body and hand lengths. Therefore, we want to test whether body length and hand length are significantly *correlated* in men. Follow the four-step process in your analysis, as shown in Example 23.4.

CONFIDENCE INTERVALS FOR THE REGRESSION SLOPE

The slope ββ of the population regression line is usually the most important parameter in a regression problem. The slope is the rate of change of the mean response as the explanatory variable increases. We often want to estimate β.β. The slope bb of the least-squares line is an unbiased estimator of β.β. A confidence interval is more useful because it shows how accurate the estimate bb is likely to be. The confidence interval for ββ has the familiar form

estimate ± t*SEestimate

$$\text{estimate} \pm t * \text{SE}_{\text{estimate}}$$

Because bb is our estimate, the confidence interval is b±t*SEb.$b \pm t^{*}\text{SE}_{b}$. Here are the details.

CONFIDENCE INTERVAL FOR REGRESSION SLOPE

A level CC **confidence interval for the slope** ββ of the population regression line is

b±t*SEb

$$b \pm t^{*}\text{SE}_{b}$$

Here t*t^{*} is the critical value for the t(n-2)$t\,(n-2)$ density curve with area CC between -t*$-t^{*}$ and t*.t^{*}. Most technology with basic statistical inference capabilities will perform all necessary computations and produce the lower and upper boundaries of the confidence interval.

EXAMPLE 23.5 Aging and the blood–brain barrier: estimating the slope

The outputs in Figure 23.2 give the slope b=0.00909$b = 0.00909$ and its standard error SEb=0.00220.$\text{SE}_{b} = 0.00220$. Some of them (Excel, JMP, and Minitab) also include the lower and upper endpoints of the 95%

> confidence interval for the population slope β,$\beta,$ 0.00452 and 0.01365. We are 95% confident that, in the hippocampus of healthy adults, mean BBB permeability increases by between about 0.00452 and 0.01365 10^{-3}/min for each additional year of age.

The computations performed by software in this example involve the following steps:

- Slope: b=0.00909$b = 0.00909$
- Standard error of the slope: SEb=0.00220$\text{SE}_b = 0.00220$
- Degrees of freedom of the *t* distribution: n-2=24-2=22$n - 2 = 24 - 2 = 22$
- Critical value for a 95% confidence level under t(22): t*=2.074 $t\,(22): \; t^* = 2.074$ (rounded to three decimal places, obtained from software or Table C)
- Margin of error: m=t*SEb=2.074(0.00220)=0.00456 $m = t^*\text{SE}_b = 2.074\,(0.00220) = 0.00456$
- 95% confidence interval for β:b±m=0.00909±0.00456, $\beta: b \pm m = 0.00909 \pm 0.00456,$ or 0.00453$0.00453$ to 0.01365$0.01365$

This agrees, up to rounding error, with the ouputs from Excel, JMP, and Minitab in Figure 23.2. Likewise, you can find a confidence interval for the intercept αα of the population regression line, using aa and SEaSE_a from the "Intercept" (or "Constant") line in most software outputs. Excel, JMP, and Minitab already include this interval in their outputs. However, we rarely need to estimate α.$\alpha.$

APPLY YOUR KNOWLEDGE

23.7 Algal bloom control: estimating slope. Exercise 23.1 gives data on the number of *D. magna* grazers introduced in test tubes and the net growth rate of *G. semen* algae. Software tells us that the least-squares slope is b=-0.5286$b = -0.5286$ with standard error SEb=0.1059.$\text{SE}_b = 0.1059.$ Give a 95% confidence interval for the population slope ββ and interpret it in context.

23.8 Frog mating calls: estimating slope. Exercise 23.2 gives data on the mating call frequency of *H. chrysoscelis* as a function of habitat temperature. Use the Minitab output in Figure 23.4 to obtain a 95% confidence interval for the slope of the population regression line. Say in words what the slope of the population regression line tells us about the frequency of this species’ mating call in the wild under various temperatures.

INFERENCE ABOUT PREDICTION

One of the most common reasons to fit a line to data is to predict the response to a particular value of the explanatory variable. This is another setting for regression inference: We want not simply a prediction, but a prediction with a margin of error that describes how accurate the prediction is likely to be.

EXAMPLE 23.6 Beer and blood alcohol

STATE: The EESEE story "Blood Alcohol Content" on the companion website describes a study in which 16 student volunteers of legal drinking age from Ohio State University drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood (g/dl). Here are the data:

Student	1	2	3	4	5	6	7	8
Beers	5	2	9	8	3	7	3	5
BAC	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06
Student	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Beers	3	5	4	6	5	7	1	4
BAC	0.02	0.05	0.07	0.10	0.085	0.09	0.01	0.05

Bildagentur-online/Schoening/Alamy Stock Photo

The students were equally divided between men and women and varied in weight and usual drinking habits, so we may be willing to consider them a representative sample of the population of young adults. What can we predict, based on this study, about the BAC of a young adult 30 minutes after finishing 5 cans of beer?

PLAN: Regress BAC on number of beers. Use the regression line to predict BAC. Give a margin of error that allows us to have 95% confidence in our prediction.

SOLVE: The scatterplot in Figure 23.7 and the regression output in Figure 23.8 show that the number of beers drunk predicts blood alcohol content quite well. Because r2=0.80,$r^2 = 0.80$, number of beers drunk explains 80% of the observed variation in BAC. To predict BAC after 5 beers, use the equation of the regression line:

y^=-0.0127+0.0180x=-0.0127+0.0180(5)=0.077

$$\begin{aligned}\hat{y} &= -0.0127 + 0.0180x \\ &= -0.0127 + 0.0180\,(5) = 0.077\end{aligned}$$

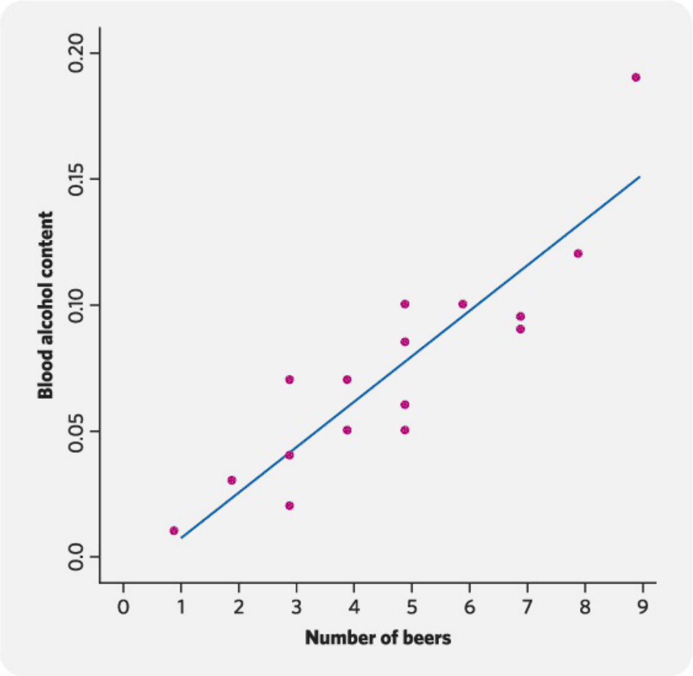
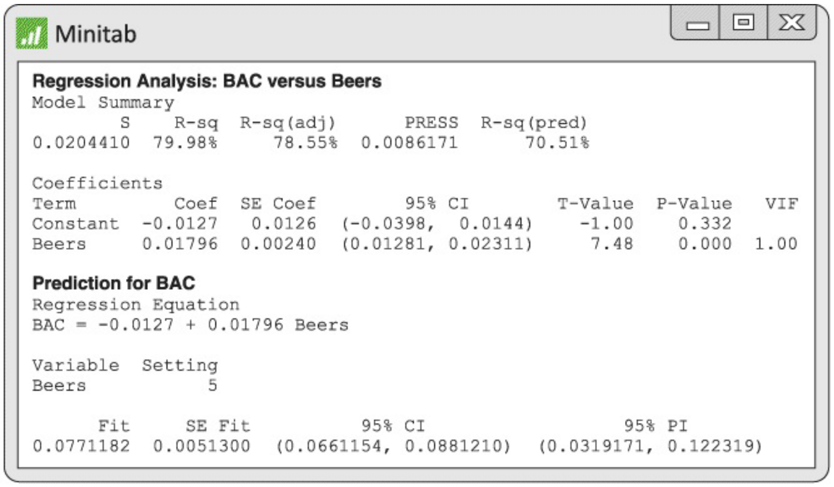


Figure 23.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 23.7 Scatterplot of college students’ blood alcohol content (BAC) against the number of cans of beer consumed, with the least-squares regression line.



Regression Analysis: BAC versus Beers

```
Model Summary
        S     R-sq  R-sq(adj)       PRESS  R-sq(pred)
0.0204410   79.98%     78.55%   0.0086171      70.51%

Coefficients
Term           Coef  SE Coef          95% CI          T-Value  P-Value   VIF
Constant   -0.0127   0.0126   (-0.0398,  0.0144)       -1.00    0.332
Beers      0.01796  0.00240   (0.01281, 0.02311)        7.48    0.000  1.00
```

Prediction for BAC

```
Regression Equation
BAC = -0.0127 + 0.01796 Beers

Variable  Setting
Beers           5

      Fit     SE Fit            95% CI                      95% PI
0.0771182  0.0051300  (0.0661154, 0.0881210)  (0.0319171, 0.122319)
```

Figure 23.8

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 23.8 Minitab output for the blood alcohol content data, with prediction for x*=5.

$$x^* = 5.$$

The legal limit for driving is a BAC of 0.08 in all states, for individuals of legal drinking age. The prediction 0.077 is very close to this legal limit 0.08. What about 95% confidence? The “Prediction for BAC” part of the output in Figure 23.8 shows *two* 95% intervals. Which should we use?

To decide which interval to use, you must answer this question: Do you want to predict the *mean* BAC for *all young adults* who drink 5 cans of beer, or do you want to predict the BAC of *one individual young adult* who drinks 5 beers? *Both of these predictions may be interesting, but they are two different problems.* The actual, sample-based prediction is the same: y^=0.077$\hat{y} = 0.077$ (“Fit” in the Minitab output). The margin of error, however, is different for the two kinds of prediction. Individual young adults who drink 5 beers don’t all have the same BAC. So we need a larger margin of error to pin down one person’s result than we would to estimate the mean BAC for all young adults who would drink 5 beers.

Write the given value of the explanatory variable xx as x*.x^*. In Example 23.6, x*=5.$x^* = 5$. The distinction between predicting a single outcome and predicting the mean of all outcomes when x=x*$x = x^*$ determines what margin of error is correct. To emphasize the distinction, we use different terms for the two intervals.

- To estimate the *mean* response, we use a *confidence interval*. It is an ordinary confidence interval for the mean response when xx has the value x*,x^*, which is μy=α+βx*.$\mu_y = \alpha + \beta x^*$. This is a parameter, a fixed number whose value we simply don’t know.
- To estimate an *individual* response y,y, we use a **prediction interval.** A prediction interval estimates a single random response yy rather than a parameter like μy.μ_y. The response yy is not a fixed number. If we took many observations with x=x*,$x = x^*$, we would get many different responses.

EXAMPLE 23.7 Beer and blood alcohol: conclusion

The output in Figure 23.8 helpfully labels the confidence interval as “CI” and the prediction interval as “PI.” The confidence interval simply says that we can be 95% confident that the *mean* BAC of all young adults who drink 5 beers is somewhere between 0.066 and 0.088 g/dl 30 minutes after finishing the drinks. To predict BAC for one individual, however, we must use the much wider prediction interval, 0.032 to 0.122 g/dl. That is, we are 95% confident that the BAC of the next young adult who drinks 5 beers will be between 0.032 and 0.122 g/dl 30 minutes after finishing the drinks. The upper part of that range exceeds the 0.08 g/dl legal limit.

The meaning of a prediction interval is very much like the meaning of a confidence interval. A 95% prediction interval, like a 95% confidence interval, is right 95% of the time in repeated use. “Repeated use” in this context means that we take an observation on yy for each of the nn values of xx in the original data, and then take one more observation yy with x=x*.$x = x^*$. We form the prediction interval from the nn observations, then see if it covers the one more y.y. In 95% of all repetitions, it will do so.

The precise interpretation of prediction intervals is a minor point. The main point is that it is harder to predict one response than to predict a mean response. Both intervals have the usual form

y^±t*SE

$$\hat{y} \pm t^{*}\,\mathrm{SE}$$

but the prediction interval is wider than the confidence interval because individuals are more variable than averages. You will rarely need to know the details, because software automates the calculation, but here they are.

CONFIDENCE AND PREDICTION INTERVALS FOR REGRESSION RESPONSE

A level CC **confidence interval for the mean response** μyμ_y when xx takes the value x*x^* is

y^±t*SEμ^

$$\hat{y} \pm t^{*}\mathrm{SE}_{\hat{\mu}}$$

The standard error SEμ^$\mathrm{SE}_{\hat{\mu}}$ is

SEμ^=s1n+(x*-x¯)2∑(x-x¯)2

$$\mathrm{SE}_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^{*}-\bar{x})^{2}}{\sum(x-\bar{x})^{2}}}$$

A level CC **prediction interval for a single observation** yy when xx takes the value x*x^* is

y^±t*SEy^

$$\hat{y} \pm t^{*}\mathrm{SE}_{\hat{y}}$$

The standard error for prediction SEy^$\mathrm{SE}_{\hat{y}}$ is

SEy^=s1+1n+(x*-x¯)2∑(x-x¯)2

$$\mathrm{SE}_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^{*}-\bar{x})^{2}}{\sum(x-\bar{x})^{2}}}$$

In both intervals, t*t^* is the critical value for the t(n-2)$t\,(n-2)$ density curve with area CC between -t*$-t^*$ and t*.t^*. Most comprehensive statistical programs will perform all necessary computations and produce the lower and upper boundaries of these intervals.

There are two standard errors: SEμ^$\mathrm{SE}_{\hat{\mu}}$ for estimating the mean response μy,μ_y, and SEy^$\mathrm{SE}_{\hat{y}}$ for predicting an individual response y.y. The only difference between the two standard errors is the extra 1 under the square root sign in the standard error for prediction. The extra 1 makes the prediction interval wider. Both standard errors are multiples of the regression standard error s.s. The degrees of freedom are again n-2,$n-2$, the degrees of freedom of s.s.

STATS IN YOUR WORLD

IS REGRESSION GARBAGE?

No—but garbage can be the setting for regression. The U.S. Census Bureau once asked if weighing a neighborhood's garbage would help count its people. So 63 households had their garbage sorted and weighed. It turned out that pounds of plastic in the trash gave the best garbage prediction of the number of people in a neighborhood. The margin of error for a 95% prediction interval in a neighborhood of about 100 households, based on five weeks' worth of garbage, was about ±2.5± 2.5 people. Alas, that was not accurate enough to help the Census Bureau.

APPLY YOUR KNOWLEDGE

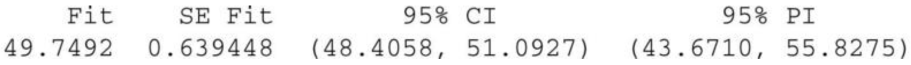
23.9 Algal bloom control: prediction. Exercise 23.1 gives data on the number of *D. magna* grazers introduced in test tubes and the net growth rate of *G. semen* algae. Analysis of the data shows a linear relationship between these two variables. Here is the partial Minitab output for prediction when x*=3:$x^* = 3$:

```
     Fit    SE Fit          95% CI                 95% PI
-2.88571 0.188532 (-3.40916, -2.36227) (-4.22295, -1.54848)
```

a. Use this output to predict the net algal growth rate in this test tube if researchers were to set up another experiment with 3 grazers in the test tube. Which of the two intervals did you use for your prediction and why?

b. Minitab provides only one of the two standard errors used in prediction: SEμ^$SE_{\hat{\mu}}$ ("SE Fit"), the standard error for estimating the mean response. Use this fact to show how Minitab computed the 95% confidence interval for the mean net algal growth rate for all test tubes containing 3 grazers.

23.10 Frog mating calls: prediction. Analysis of the data in Exercise 23.2 shows that mating-call frequency in *H. chrysoscelis* varies linearly with habitat temperature. We might want to predict the species' mean mating-call frequency (in notes per second) at 24 degrees Celsius in the wild. Here is the partial Minitab output for prediction when x*=24$x^* = 24$ degrees Celsius:



```
    Fit     SE Fit          95% CI                 95% PI
49.7492   0.639448   (48.4058, 51.0927)   (43.6710, 55.8275)
```

a. Use the regression line from Figure 23.4 to verify that "Fit" is the predicted value for x*=24$x^* = 24$ (up to roundoff error).

b. What is the 95% interval we want? Interpret this interval in context.

CHECKING THE CONDITIONS FOR INFERENCE

You can fit a least-squares line to any set of explanatory-response data when both variables are quantitative and at least the response variable is continuous. If the scatterplot doesn't show a roughly linear pattern, the fitted line may be useless or, worse, misleading. But it is still the line that fits the data best in the least-squares sense. To use regression inference, however, the data must satisfy additional conditions. *Before you can trust the results of inference, you must check the conditions for inference one by one.* There are ways to deal with violations of any of the conditions. If you see a clear violation, get expert advice.

CAUTION

Although the conditions for regression inference are a bit elaborate, it is not hard to check for major violations. The conditions involve the population regression line and the deviations of responses from this line. We can't observe the population line, but the least-squares line estimates it, and the residuals estimate the deviations from the population line. *You can check all the conditions for regression inference by looking at graphs of the residuals.* Most regression software will calculate and save the residuals for you, and even plot them. Start by examining a dotplot or histogram of the residuals and also a **residual plot,** a plot of the residuals against the explanatory variable x,x, with a horizontal line at the “residual=0”“$\text{residual} = 0$” position. The “residual=0”“$\text{residual} = 0$” line represents the position of the least-squares line in the scatterplot of yy against x.x. Let's look at each condition in turn.

- **The relationship is linear in the population.** Look for curved patterns or other departures from a straight-line overall pattern in the residual plot. You can also use the original scatterplot, but the residual plot magnifies any effects.
- **The response varies Normally about the population regression line.** Because different *y*-values usually come from different *x*-values, the responses themselves need not be Normal. It is the deviations from the population line—estimated by the residuals—that must be Normal. Check for clear skewness or other major departures from Normality in a histogram of the residuals or by making a Normal quantile plot of the

residuals.

- **Observations are independent.** In particular, repeated observations on the same individual are not allowed. You should not use ordinary regression to make inferences about the growth of children measured more than once over time, for example. Signs of dependence in the residual plot are a bit subtle, so we usually rely on common sense and information from the data collection process.
- **The standard deviation of the responses is the same for all values of** x.x. Look at the scatter of the residuals above and below the “residual=0”$\text{“residual} = 0\text{”}$ line in the residual plot. The scatter should be roughly the same from one end to the other. You will sometimes find that as the response yy gets larger, so does the scatter of the residuals. Rather than remaining fixed, the standard deviation σσ about the line changes with xx as the mean response changes with x.x. There is no fixed σσ for ss to estimate. You cannot trust the results of inference when this happens.

You will always see some irregularity when you look for Normality and fixed standard deviation in the residuals, especially when you have few observations. Don’t overreact to minor violations of the conditions. Like other tt procedures, inference for regression is (with one exception) not very sensitive to lack of Normality, especially when we have many observations. Do beware of influential observations, which can greatly affect the results of inference.

The exception is the prediction interval for a single response y.y. This interval relies on Normality of individual observations, not just on the approximate Normality of statistics like the slope a a and intercept bb of the least-squares line. The statistics aa and bb become more Normal as we take more observations. This contributes to the robustness of regression inference, but it isn’t enough for the prediction interval. We will not study methods that carefully check Normality of the residuals, so *you should regard prediction intervals as rough approximations.*

EXAMPLE 23.8 Aging and the blood–brain barrier: checking

conditions

Before we can trust the conclusions from the inference we performed in Examples 23.3 and 23.4, we need to check the conditions for regression inference.

- **Linear relationship.** The scatterplot in Figure 23.1 shows a linear relationship with possibly one high outlier. The residual plot in Figure 23.9(a) shows no unusual pattern that would indicate curvature.
- **Normal residuals.** The histogram of residuals in Figure 23.9(b) is approximately symmetric and single-peaked. There are no clear departures from Normality: There are some extreme observations but they appear on both sides of the distribution peak.
- **Independent observations.** The 24 healthy adults recruited for the study were unrelated, so we are willing to regard them as independent observations.
- **Constant standard deviation.** While we notice a lot of variation on either side of the “residual=0”“residual $= 0$” line in the residual plot, there is no unusual variation in the scatter of the residuals above and below the line as xx varies. One observation, however, has a particularly large residual.

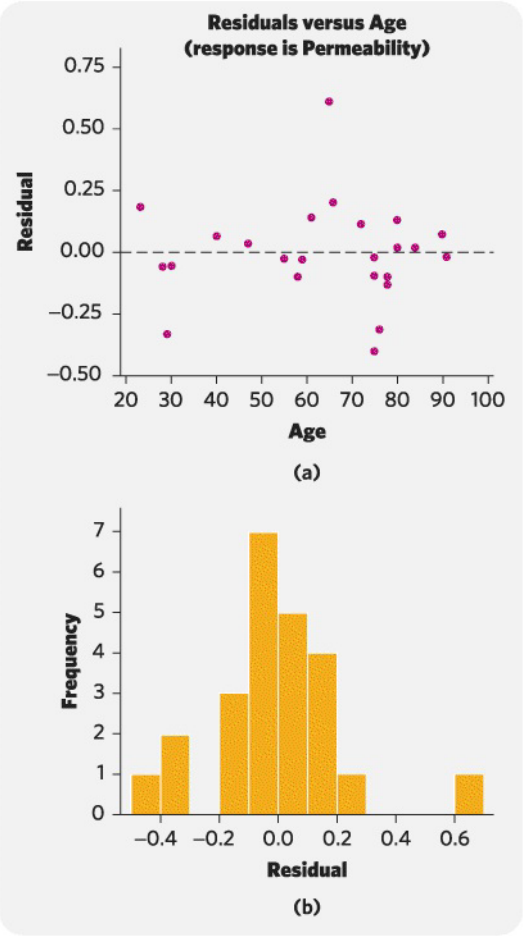


Figure 23.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

FIGURE 23.9 Residual plot (a) and histogram of residuals (b) for the regression of BBB permeability on age.

The outlier (age 65 and permeability 1.85, shown as an open circle in Figure 23.10) is the only potentially serious violation of the conditions for inference. How influential is this outlier? The dashed line in Figure 23.10 is the regression line without the outlier. The two regression lines have nearly the same slope (0.00909 with all 24 points compared with 0.00896 when excluding the outlier). That's because x=65$x = 65$ is near the mean of xx and there are many other observations with a similar age around the outlier in the data set. Exercise 23.38 explores how influential a point can be if it is an extreme value of x.x.

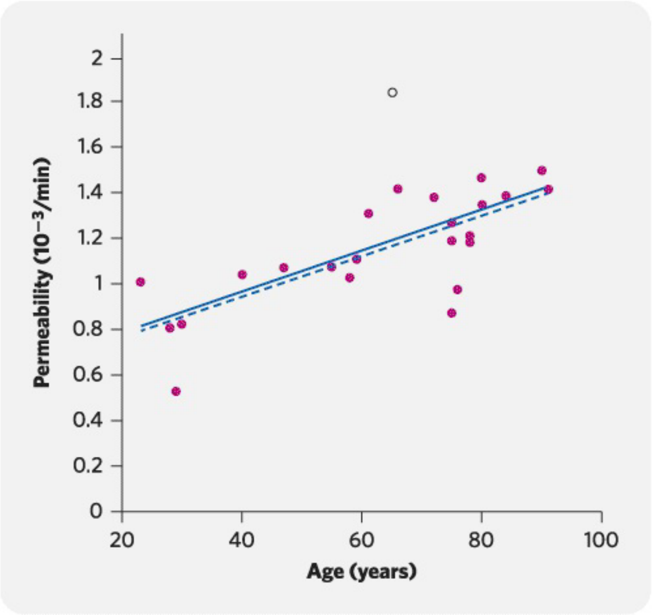


Figure 23.10
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 23.10 Scatterplot of blood-brain barrier permeability in the hippocampus of 24 healthy adults as a function of their age. The two regression lines are for the data with (solid line) and without (dashed line) the unusual observation (open circle).

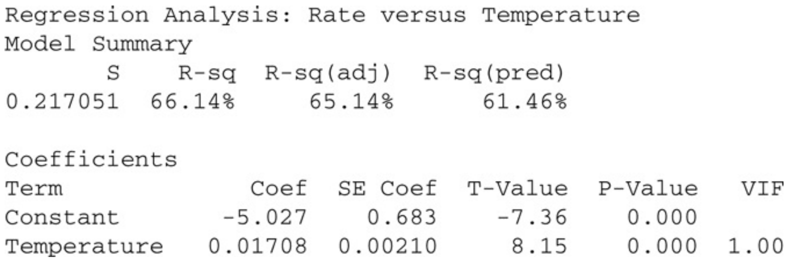
Even though the outlier in Example 23.8 is not very influential for the slope of the regression line, it influences regression inference because of its effect on the regression standard error. Without the outlier, the overall amount of scatter is decreased, lowering the regression standard error from 0.216 to 0.168. While this would lead to an even smaller *P*-value or a narrower confidence interval about the population slope, it would not fundamentally alter our main conclusion from the study because the value of the slope is not substantially affected.

Of course, *whether or not an outlier is influential should have no bearing on whether we should keep it when analyzing the data.* We first considered the treatment of outliers in Chapter 2 along with an in-depth discussion on page 55. In Example 3.6 (page 73) we also examined the treatment of outliers specifically in the context of correlation and regression. In Example 23.8, legitimate reasons for excluding the unusual observation would be if it turned out to be a data entry error or if further medical examination indicated that the subject was not actually a healthy adult. This was not the case, however: The researchers did not consider the outlier to be a physiologically suspicious numerical value and they very carefully examined both the physical and mental health of all subjects.

One more caution about inference in this example: As usual in an observational study, the possibility of lurking variables makes us hesitant to conclude that increasing age is *causing* the observed increased BBB permeability in the hippocampus of healthy adults. The effect of any lurking variable that increased over time—perhaps a cumulative lifetime exposure to environmental toxins or a gradual reduction in exercise—would be confounded with the effect of age. This is not a trivial distinction: Age cannot be changed at will, but pollution and lack of exercise can be addressed.

APPLY YOUR KNOWLEDGE

23.11 Enzyme activity and temperature. Exercise 3.4 (page 74) gives data on the activity rate of the digestive enzyme acid phosphatase in vitro at varying temperatures. Is there significant evidence of straight-line dependence between temperature and acid phosphatase activity rate? Here is Minitab's least-squares regression output for the data:



```
Regression Analysis: Rate versus Temperature
Model Summary
       S     R-sq  R-sq(adj)  R-sq(pred)
0.217051  66.14%     65.14%      61.46%

Coefficients
Term            Coef  SE Coef  T-Value  P-Value   VIF
Constant      -5.027    0.683    -7.36    0.000
Temperature  0.01708  0.00210     8.15    0.000  1.00
```

a. If you haven't already, display the data in a scatterplot and describe the pattern.

b. What are the null and alternative hypotheses about the population slope tested by Minitab? What is the reported *P*-value for this test?

c. Use these data to explain how a linear regression analysis can give a statistically significant *P*-value in the absence of a linear pattern, and why it is important to *always* plot the data and check the conditions when performing statistical inference.

23.12 Frog mating calls: residuals. Figure 23.4 gives part of the Minitab output for the data on mating-call frequency and temperature in wild *H. chrysoscelis* from Exercise 23.2. Table 23.2 comes from another part of the output. It gives the predicted response y^$\hat{y}$ and the residual y-y^$y - \hat{y}$ for each of the 20 observations. Most statistical software packages provide similar output. Examine the conditions for regression inference one by one.

TABLE 23.2 Frog mating calls: predictions and residuals

Obs.	Temp. xx	Freq. yy	Prediction y^$\hat{y}$	Residual y-y^$y - \hat{y}$
1	19	38	38.10	–0.10
2	21	42	42.76	–0.76
3	22	45	45.09	–0.09
4	22	45	45.09	–0.09
5	23	41	47.42	–6.42
6	23	45	47.42	–2.42
7	23	48	47.42	0.58
8	23	50	47.42	2.58
9	23	53	47.42	5.58
10	24	51	49.75	1.25
11	24	48	49.75	–1.75
12	24	53	49.75	3.25
13	24	47	49.75	–2.75
14	25	53	52.08	0.92
15	25	49	52.08	–3.08
16	25	56	52.08	3.92
17	25	53	52.08	0.92
18	26	55	54.41	0.59
19	26	55	54.41	0.59
20	27	54	56.74	–2.74

a. **Linear relationship.** Your plot and r2r^2 from Exercise 23.2 show that the relationship is indeed linear. Residual plots magnify effects. Plot the residuals against temperature. Are there clear deviations from a linear relationship?

b. **Normal variation about the line.** Make a histogram of the residuals. Do strong skewness or outliers suggest lack of Normality?

c. **Independent observations.** The data come from observations of 20 different male *H. chrysoscelis* in the wild. They are therefore independent observations.

d. **Spread about the line stays the same.** Your plot in part b shows that the spread about the line does not stay the same: Variation is greater for warmer temperatures. However, the reason for less variation at lower temperatures is that there are very few observations. Rather than a problem with constant spread, this suggests that the low-temperature observations might be more influential.

CHAPTER 23 SUMMARY

- **Least-squares regression** fits a straight line to data to predict a quantitative response variable yy from a quantitative explanatory variable x.x. Inference about regression requires more conditions.
- The **conditions for regression inference** say that there is a **population regression line** μy=α+βx$\mu_y = \alpha + \beta x$ that describes how the mean response varies as xx changes. The distribution of observed responses y y for any xx is Normal with mean given by the population regression line and with the same standard deviation σσ for any value of x.x. Observations on yy are independent.
- The **parameters to be estimated** are the intercept αα and the slope ββ of the population regression line as well as the regression standard deviation σ.σ. The intercept aa and the slope bb of the least-squares line estimate αα and β,β, respectively. The **regression standard error** ss estimates σ.σ.
- The regression standard error ss has n-2$n - 2$ **degrees of freedom.** All *t* procedures in regression inference have n-2$n - 2$ degrees of freedom, and statistical software should be used to automate the computations whenever possible.
- Testing **the hypothesis that the slope is zero in the population** relies on the *t* statistic t=b/SEb.$t = b/\text{SE}_b$. This null hypothesis says that straight-line dependence on xx is not relevant for predicting y.y.
- The *t* test for regression slope is also a test for **the hypothesis that the population correlation between** xx **and** yy **is zero.** Table E at the back of the book can be used to do this test with just the sample correlation r. r.
- **Confidence intervals for the slope** of the population regression line have the form b±t*SEb.$b \pm t^* \text{SE}_b$.
- **Confidence intervals for the mean response** when xx has value x*x^* have the form y^±t*SEμ^.$\hat{y} \pm t^* \text{SE}_{\hat{\mu}}$. **Prediction intervals** for an individual future response yy have a similar form with a larger standard

error, y^±t*SEy^.$\hat{y} \pm t^{*}\mathrm{SE}_{\hat{y}}$. For a given value x*$x^{*}$ of x,x, the prediction interval for a future response is always larger than the confidence interval for the mean response. In addition, unlike confidence intervals for a mean response, prediction intervals are not robust to deviations from Normality in the distribution of residuals.

THIS CHAPTER IN CONTEXT

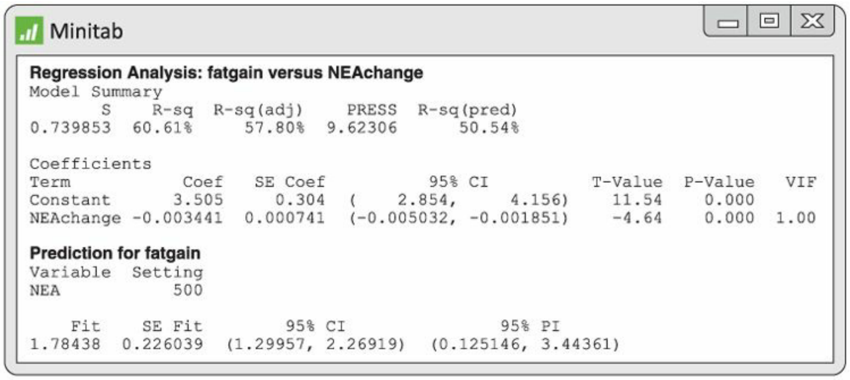
In Chapters 3 and 4 we discussed how to describe the relationship between *two quantitative variables*. We used scatterplots to visualize patterns and identify linear relationships in particular. We then quantified the strength and direction of a linear relationship with the correlation coefficient r r and obtained the equation of the least-squares regression line to model that relationship. In this chapter we expand these concepts beyond the domain of descriptive statistics into that of statistical inference to ask whether the linear pattern in the sample data would hold for the entire population.

The statistical inference methods described in this chapter draw on *t* procedures similar to those seen in Chapter 17. We can use them to estimate the parameters of the regression line or to test the hypothesis of no association (a slope of zero) between the two variables. Even if the association is statistically significant, we cannot necessarily conclude that there is an underlying cause-and-effect relation between the response and explanatory variables. As we saw in Chapters 6 and 7, the way the data were collected can substantially limit the extent of our conclusions.

In companion Chapter 28 we expand the concept of simple linear regression to include multiple explanatory variables used to describe and predict one response variable (multiple linear regression). We use inference procedures to determine which of these explanatory variables, alone or in sets, has a statistically significant effect on the response variable. We also see how a categorical variable can be coded with zeros and ones to allow it into the regression model (in multiple linear regression and in logistic regression).

CHECK YOUR SKILLS

Our first example of regression (Example 4.1, page 94) presented data showing that people who increased their nonexercise activity (NEA) when they were deliberately overfed gained less fat than other people. The scatterplot and regression line based on 16 overfed subjects are displayed in Figure 4.1, page 94. Figure 23.11 shows the Minitab output for regressing fat gain on NEA change in this study, along with prediction for a person adding 500 NEA calories. Exercises 23.13 to 23.22 are based on this information.



Regression Analysis: fatgain versus NEAchange

```
Model  Summary
        S     R-sq  R-sq(adj)     PRESS  R-sq(pred)
0.739853  60.61%      57.80%  9.62306      50.54%

Coefficients
Term             Coef   SE Coef            95% CI            T-Value  P-Value   VIF
Constant        3.505     0.304  (     2.854,      4.156)      11.54    0.000
NEAchange -0.003441  0.000741  (-0.005032, -0.001851)     -4.64    0.000  1.00
```

Prediction for fatgain

```
Variable  Setting
NEA           500

    Fit    SE Fit         95% CI                  95% PI
1.78438  0.226039  (1.29957, 2.26919)  (0.125146, 3.44361)
```

Figure 23.11

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 23.11 Minitab output for the nonexercise activity (NEA) change and fat gain data, with prediction for x*=500.

$$x^* = 500.$$

23.13 The equation of the least-squares regression line for predicting fat gain from NEA change is

a. fatgain=11.54-4.64 NEAchange.$\text{fatgain} = 11.54 - 4.64\,\text{NEAchange}.$

b. fatgain=-0.0034+3.505 NEAchange.$\text{fatgain} = -0.0034 + 3.505\,\text{NEAchange}.$

c. fatgain=3.5051-0.0034 NEAchange.$\text{fatgain} = 3.5051 - 0.0034\,\text{NEAchange}.$

23.14 What is the correlation between fat gain and NEA change?

a. 0.606

b. 0.778
c. − 0.778

23.15 What is the regression standard error for these data?

a. 0.0007
b. 0.3036
c. 0.7399

23.16 Is there significant evidence that fat gain decreases as NEA change increases? To answer this question, we test the hypotheses

a. H0:β=0$H_0 : \beta = 0$ versus Ha:β<0.$H_a : \beta < 0.$
b. H0:β=0$H_0 : \beta = 0$ versus Ha:β≠0.$H_a : \beta \neq 0.$
c. H0:α=0$H_0 : \alpha = 0$ versus Ha:α<0.$H_a : \alpha < 0.$

23.17 Based on the Minitab output, we conclude that

a. there is very strong evidence (P<0.0005)$(P < 0.0005)$ that fat gain decreases as NEA change increases in the population of overfed individuals.
b. there is significant evidence that fat gain decreases a lot (P<0.0005)$(P < 0.0005)$ when NEA change increases in the population of overfed individuals.
c. the study failed to find statistically significant evidence (P=0.7399)$(P = 0.7399)$ of a linear association between NEA change and fat gain in the population of overfed individuals.

23.18 Which of the intervals in the Minitab output is the 95% confidence interval for the population slope β?β?

a. (2.854, 4.156)
b. (−0.005032, −0.001851)
c. (1.29957, 2.26919)

23.19 The interpretation in context of the 95% confidence interval for the population slope ββ is that we are 95% confident that

a. fat gain decreases by some amount in this interval for every 1 additional Calorie increase in NEA.
b. fat gain decreases, on average, by some amount in this interval for

every 1 additional Calorie increase in NEA.

c. 95% of individuals who increase their NEA by 1 Calorie decrease their fat gain by some amount in this interval.

23.20 The Minitab output includes a prediction for y y when $x^*=500$. $x^* = 500$. If an overfed adult burned an additional 500 NEA calories, we can be 95% confident that the person’s fat gain would be between

a. 0.13 and 3.44 kg.
b. 1.30 and 2.27 kg.
c. 2.85 and 4.16 kg.

23.21 If a whole population of overfed adults burned an additional 500 NEA calories, we can be 95% confident that the population mean fat gain would be between

a. 0.13 and 3.44 kg.
b. 1.30 and 2.27 kg.
c. 2.85 and 4.16 kg.

23.22 To check the conditions for inference for linear regression we should in particular

a. confirm that the *P*-value is less than the significance level.
b. examine the distribution of the explanatory variable NEAchange in a histogram or a dotplot.
c. examine the distribution of the residuals in a histogram or a residual plot.

CHAPTER 23 EXERCISES

23.23 Too much nitrogen? Intensive agriculture and burning of fossil fuels increase the amount of nitrogen deposited on the land. Too much nitrogen can reduce the variety of plants by favoring rapid growth of some species—think of putting fertilizer on your lawn to help grass choke out weeds. A study of 68 grassland sites in Britain measured nitrogen deposited (kilograms of nitrogen per hectare of land area per year) and assessed the "richness" of plant species (based on number of species and how abundant each species is). The authors reported a regression analysis as follows:[6]

$$\text{plant species richness}=23.3-0.408\times\text{nitrogen deposited}$$

$$\text{plant species richness} = 23.3 - 0.408 \times \text{nitrogen deposited}$$

$$r2=0.55 \quad P<0.0001$$

$$r^2 = 0.55 \; P < 0.0001$$

a. What does the slope $b=-0.408$$b = -0.408$ say about the effect of increased nitrogen deposits on species richness?

b. What does $r2=0.55$$r^2 = 0.55$ add to the information given by the equation of the least-squares line?

c. What null and alternative hypotheses do you think the *P*-value refers to? What does this *P*-value tell you?

23.24 Heritability of phenotype. Figure 23.12 shows the relationship between parent nest size and offspring nest size for 100 parent–offspring observations of barn swallows, *Hirundo rustica*.[7] The line is the least-squares regression line. Note that the graph actually plots the logarithm of nest size, because nest sizes, like many measures of volume, are otherwise strongly skewed to the right. Here is a partial Excel regression output for these data:

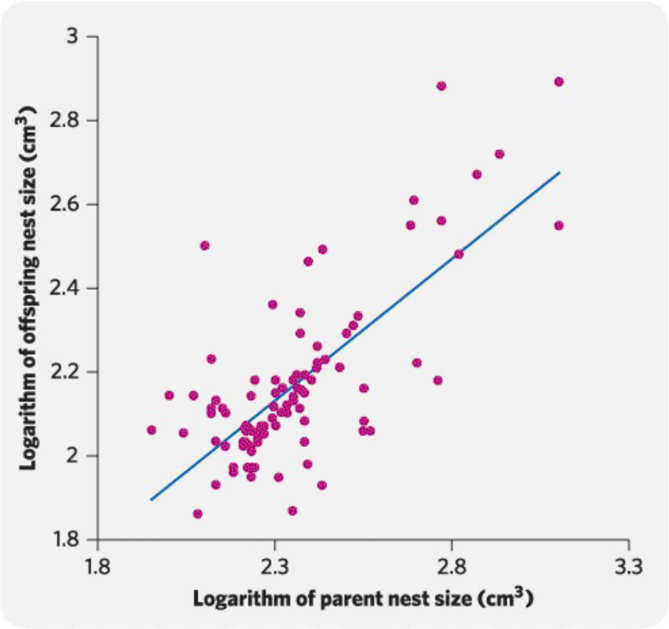


Figure 23.12
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 23.12 Scatterplot of the log of offspring nest size against the log of parent nest size.

```
Regression Statistics
Multiple R        0.7229
R Square          0.5226
Standard Error    0.1396
Observations      100

          Coefficients  Standard Error  t Stat  P-value
Intercept       0.5764          0.1542
Parents         0.6762          0.0653
```

a. Describe the relationship from the graph and assess the validity of inference for these data.

b. Give the value of r2r^2 and the equation of the least-squares line.

c. Based on the information given, test the hypothesis that there is no straight-line relationship between parent and offspring nest size. Give a test statistic and its *P*-value (using technology or Table C), then conclude in context.

23.25 The amygdala and memory. The amygdala is a brain structure involved in the processing and memory of emotional reactions. In a research project, 10 subjects were shown emotional video clips. Then they had their brains scanned with positron emission tomography (PET), and their memory of the clips was assessed quantitatively. Here are the relative amygdala activity and the memory score for each subject:[8]

Relative activity	–0.417	–0.258	–0.234	–0.249	–0.156
Memory score	31	29	29	30	33
Relative activity	–0.031	0.120	0.240	0.342	0.654
Memory score	32	31	35	34	33

a. Make a scatterplot that shows the relationship between relative amygdala activity (explanatory) and memory score (response).

b. The scatterplot shows a positive linear relationship. How strong is it? Compute r.r.

c. With only 10 data points, we wonder if the relationship is statistically significant. Run the appropriate test, find the *P*-value, and conclude in context. The standard error of the least-squares slope is SEb=1.607. $\mathrm{SE}_b = 1.607$.

23.26 Heritability of phenotype: estimating the slope. Regression is used in genetics to assess the heritability of a given phenotype from parents to their offspring. An estimate of heritability can be obtained by plotting the phenotypic value of a sample of offspring against that of their parents and finding the slope of the corresponding regression line. Using the information from Exercise 23.24, give a 95% confidence interval for the true population heritability (slope ββ) of nest size on a logarithmic scale.

23.27 The amygdala and memory: estimating the slope. Refer to Exercise 23.25. Find the 95% confidence interval for the slope of the population regression line using either software or Table C.

Oystersare categorized for retail as small, medium, or large based on their volume. The grading process is slow and expensive when done by hand. A computer reconstruction of oyster volume based on image analysis would be desirable to speed the process. Engineers designed two programs *estimating oyster volume, one based on two-dimensional (2D) image processing and the other based on three-dimensional (3D) image processing. We want to know if either approach is a good predictor of actual oyster volume. The results of both programs are displayed in Table 23.3 for a sample of 30 oysters. Actual volumes are expressed in cubic centimeters (cm*^{3}*), 2D reconstructions in thousands of pixels, and 3D reconstruction in millions of volume pixels.*[9] *Exercises 23.28 to 23.31 are based on this information.*

TABLE 23.3 Actual volume and two computer reconstruction outputs for a sample of 30 oysters

Actual	2D	3D	Actual	2D	3D	Actual	2D	3D
13.04	47.907	5.136699	10.53	31.216	3.942783	10.95	37.156	4.707532
11.71	41.458	4.795151	10.84	41.852	4.052638	7.97	29.070	3.019077
17.42	60.891	6.453115	13.12	44.608	5.334558	7.34	24.590	2.768160
7.23	29.949	2.895239	8.48	35.343	3.527926	13.21	48.082	4.945743
10.03	41.616	3.672746	14.24	47.481	5.679636	7.83	32.118	3.138463
15.59	48.070	5.728880	11.11	40.976	4.013992	11.38	45.112	4.410797
9.94	34.717	3.987582	15.35	65.361	5.565995	11.22	37.020	4.558251
7.53	27.230	2.678423	15.44	50.910	6.303198	9.25	39.333	3.449867
12.73	52.712	5.481545	5.67	22.895	1.928109	13.75	51.351	5.609681
12.66	41.500	5.016762	8.26	34.804	3.450164	14.37	53.281	5.292105

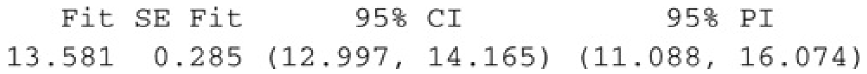
23.28 Grading oysters in 2D. Consider the original program that estimates oyster volume using 2D digital image processing.

a. Create a scatterplot of 2D volume reconstruction and actual volume, using 2D reconstruction for the explanatory variable. Find the equation for the least-squares regression line.

b. Verify the conditions for inference.

c. Test whether the linear relationship is statistically significant and give a 95% confidence interval for the slope of the population regression line.

23.29 Grading oysters in 2D: prediction. Here is a partial Minitab output for prediction when 2D reconstruction is 50 thousand pixels:



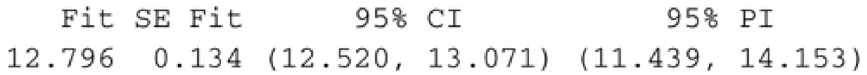
```
   Fit SE Fit      95% CI           95% PI
13.581  0.285 (12.997, 14.165) (11.088, 16.074)
```

a. Give a 95% confidence interval for the mean actual volume of oysters for which the 2D reconstruction is 50 thousand pixels.
b. Give a 95% confidence interval for the actual volume of 1 randomly selected oyster if that oyster has a 2D reconstruction of 50 thousand pixels.
c. The purpose of this computer-imaging system is to avoid grading oysters by hand. Oysters are labeled as small, medium, or large if their volume is between 0 and 9.9 cm^3, between 10 and 12.9 cm^3, or greater than 13 cm^3, respectively. Which of the two confidence intervals would you use to assign a grade label to an oyster? In your judgment, is the 2D imaging system practically useful?

23.30 Grading oysters in 3D. Now consider the newer program that estimates oyster volume using 3D digital image processing.

a. Create a scatterplot of 3D volume reconstruction and actual volume, using 3D reconstruction for the explanatory variable. Find the equation for the least-squares regression line.
b. Verify the conditions for inference.
c. Test whether the linear relationship is statistically significant and give a 95% confidence interval for the slope of the population regression line.

23.31 Grading oysters in 3D: prediction. Here is a partial Minitab output for prediction when 3D reconstruction is 5 million volume pixels:

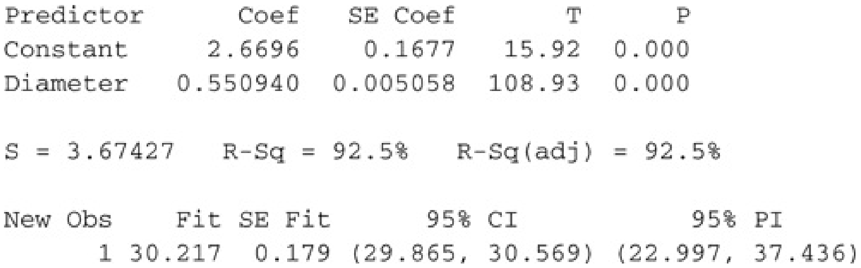


```
   Fit SE Fit      95% CI           95% PI
12.796  0.134 (12.520, 13.071) (11.439, 14.153)
```

a. Give a 95% confidence interval for the mean actual volume of oysters for which the 3D reconstruction is 5 million volume pixels.
b. Give a 95% confidence interval for the actual volume of 1 randomly selected oyster if that oyster has a 3D reconstruction of 5 million volume pixels.
c. The purpose of this computer-imaging system is to avoid grading

oysters by hand. Oysters are labeled as small, medium, or large if their volume is between 0 and 9.9 cm^3, between 10 and 12.9 cm^3, or greater than 13 cm^3, respectively. Which of the two confidence intervals would you use to assign a grade label to an oyster? In your judgment, is the 3D imaging system practically useful? How does it compare with the 2D program?

23.32 Predicting tree height. Measuring tree height is not an easy task. How well might trunk diameter predict tree height? A survey of 958 live trees in an old-growth forest in Canada answered this question.[10] Here is part of the Minitab output based on these data for regressing height on diameter, both in centimeters (cm), along with prediction for a tree having a diameter of 50 cm:



```
Predictor       Coef   SE Coef       T      P
Constant      2.6696    0.1677   15.92  0.000
Diameter    0.550940  0.005058  108.93  0.000

S = 3.67427   R-Sq = 92.5%   R-Sq(adj) = 92.5%

New Obs    Fit SE Fit       95% CI             95% PI
      1 30.217  0.179 (29.865, 30.569) (22.997, 37.436)
```

a. A scatterplot of the data shows a reasonably linear relationship between tree height and diameter. Is this relationship statistically significant? How strong is the relationship?
b. Give a 95% confidence interval for the height of 1 tree randomly selected from this forest if the tree has a diameter of 50 cm.
c. Give a 95% confidence interval for the mean height of all the trees in this forest that have a diameter of 50 cm. How does this interval compare with the one you calculated in part b?

23.33 DNA on the ocean floor. We think of DNA as the stuff that stores the genetic code. Itturns out that DNA occurs, mainly outside living cells, on the ocean floor. It is important in nourishing seafloor life. Scientists think that this DNA comes from organic matter that settles to the bottom from the top layers of the ocean. “Phytopigments,” which come mainly from algae, are a measure of the amount of organic matter that has settled to the bottom. Table 23.4 provides data on concentrations of DNA and phytopigments (both in grams per square meter) in 116 ocean locations around the world.[11] Examine the relationship between DNA and phytopigments. Do the data give good reason to think that phytopigment concentration helps explain DNA concentration? Describe the data and follow the four-step process in answering this question.

TABLE 23.4 DNA and phytopigment concentrations (g/m^2) on the ocean floor

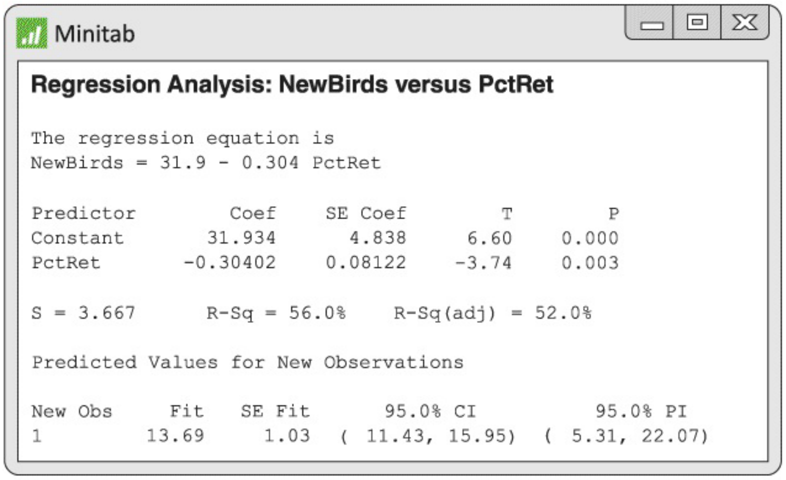
DNA	Phyto	DNA	Phyto	DNA	Phyto	DNA	Phyto	DNA	Phyto	DNA	Phyto
0.148	0.010	0.276	0.056	0.156	0.032	0.300	0.022	0.344	0.009	0.090	0.003
0.108	0.009	0.214	0.023	0.112	0.016	0.116	0.008	0.464	0.030	0.206	0.001
0.180	0.008	0.330	0.016	0.280	0.005	0.120	0.004	0.538	0.055	0.307	0.001
0.218	0.006	0.240	0.007	0.308	0.005	0.064	0.006	0.153	0.001	0.552	0.040
0.152	0.006	0.100	0.006	0.238	0.011	0.228	0.010	0.095	0	0.168	0.005
0.589	0.050	0.463	0.038	0.461	0.034	0.333	0.020	0.901	0.075	0.112	0.003
0.357	0.023	0.382	0.032	0.414	0.034	0.241	0.012	0.116	0.003	0.179	0.002
0.458	0.036	0.396	0.033	0.307	0.018	0.236	0.002	0.132	0.005	0.264	0.026
0.076	0.001	0.001	0.002	0.009	0	0.099	0	0.302	0.014	1.056	0.082
0.187	0.001	0.104	0.004	0.088	0.009	0.072	0.002	0.360	0.010	0.207	0.001
0.192	0.005	0.152	0.028	0.152	0.006	0.272	0.004	0.296	0.010	0.131	0.001
0.288	0.046	0.232	0.006	0.368	0.003	0.216	0.011	0.130	0.001	0.822	0.058
0.248	0.006	0.280	0.002	0.336	0.062	0.320	0.006	0.172	0.001	0.172	0.006
0.896	0.055	0.200	0.017	0.408	0.018	0.472	0.017	0.171	0.001	0.100	0.001
0.648	0.034	0.384	0.008	0.440	0.042	0.592	0.032	0.391	0.026	0.162	0.001
0.392	0.036	0.312	0.002	0.312	0.003	0.208	0.001	0.074	0	0.213	0.007
0.128	0.001	0.264	0.001	0.264	0.008	0.328	0.010	0.121	0.004		
0.264	0.002	0.376	0.003	0.288	0.001	0.208	0.024	0.369	0.023		
0.224	0.017	0.376	0.010	0.600	0.024	0.168	0.014	0.184	0.010		
0.264	0.018	0.152	0.010	0.184	0.016	0.312	0.017	0.328	0.028		

23.34 Sparrowhawk colonies. One of nature’s patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:[12]

Percent return xx	74	66	81	52	73	62	52	45	62	46	60	46	38
New adults yy	5	6	8	11	12	15	16	17	18	18	19	20	20

There is a moderately strong linear relationship between xx and y.$y.$ Figure 23.13 shows part of the Minitab regression output, including a prediction for yy when 60% of the previous year's adult birds return.

a. Write the equation of the least-squares line and use it to check that the "Fit" in the output is the predicted response for x*=60%.$x^* = 60\%.$
b. Which 95% interval in the output gives us a margin of error for predicting the average number of new birds in colonies to which 60% of the past year's adults return?



Regression Analysis: NewBirds versus PctRet

```
The regression equation is
NewBirds = 31.9 - 0.304 PctRet

Predictor        Coef    SE Coef       T        P
Constant       31.934      4.838    6.60    0.000
PctRet       -0.30402    0.08122   -3.74    0.003

S = 3.667       R-Sq = 56.0%     R-Sq(adj) = 52.0%

Predicted Values for New Observations

New Obs      Fit    SE Fit        95.0% CI            95.0% PI
1          13.69      1.03   ( 11.43, 15.95)   ( 5.31, 22.07)
```

Figure 23.13
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 23.13 Partial Minitab output for predicting number of new birds in a sparrowhawk colony from percent of birds returning.

23.35 DNA on the ocean floor: residuals. Save the residuals from the regression of DNA concentration on phytopigment concentration (Exercise 23.33). Examine the residuals to see how well the conditions for regression inference are met.

a. Plot the residuals against phytopigment concentration (the explanatory variable), using vertical limits −1 to 1 to make the pattern clearer. Add a horizontal line at height 0 to represent the regression line. What do you conclude about the conditions of linear relationship and constant standard deviation?

b. Make a histogram of the residuals. What do you conclude about Normality?

23.36 Sparrowhawk colonies: residuals. The regression of number of new birds that join a sparrowhawk colony on the percent of adult birds in the colony that return from the previous year is an example of data that satisfy the conditions for regression inference well. Here are the residuals for the 13 colonies in Exercise 23.34:

Percent return	74	66	81	52	73	62	52
Residual	–4.44	–5.87	0.69	–5.13	2.26	1.92	–0.13
Percent return	45	62	46	60	46	38	
Residual	–1.25	4.92	0.05	5.31	2.05	–0.38	

a. **Independent observations.** Why are the 13 observations independent?

b. **Linear relationship.** A plot of the residuals against the explanatory variable xx magnifies the deviations from the least-squares line. Does the plot show any systematic deviation from a roughly linear pattern?

c. **Constant spread about the line.** Does your plot in part b show any systematic change in spread as xx changes?

d. **Normal variation about the line.** Make a histogram of the residuals. With only 13 observations, no clear shape emerges. Do strong skewness or outliers suggest lack of Normality?

23.37 Testosterone and BMI in adolescent males. Obesity in adult males is associated with lower levels of sex hormone. A study investigated a possible link between body mass index (BMI, in kilograms per square meter, kg/m^2 and plasma testosterone concentrations (in nanomoles per liter of blood, nmol/l) in a sample of 50 adolescent males between the ages of 14 and 20 years. The data are displayed in Table 23.5.[13] Follow the four-step process in the following analysis.

a. Describe the relationship in a graph and with a regression line, using

BMI as the explanatory variable. Be sure to save the regression residuals.

b. Check the conditions for inference. Exercise 23.36 provides a handy outline.

c. Give a 95% confidence interval to estimate how plasma testosterone concentration changes as BMI increases.

TABLE 23.5 Body mass index (BMI, in kg/m^2) and testosterone concentration adolescent males

BMI	Testosterone	BMI	Testosterone	BMI	Testosterone	BMI	Testosterone
21.4	0.78	20.4	0.49	19.5	0.23	33.2	0.23
19.0	0.70	16.2	0.43	20.2	0.24	34.7	0.24
18.3	0.63	17.8	0.42	21.2	0.24	35.8	0.06
19.5	0.60	21.0	0.38	21.3	0.26	37.0	0.15
20.9	0.60	18.6	0.35	22.2	0.27	37.0	0.17
23.4	0.69	20.9	0.35	28.3	0.30	39.0	0.18
25.0	0.76	22.4	0.32	27.7	0.24	41.6	0.17
24.1	0.58	23.5	0.31	28.1	0.19	42.4	0.15
24.2	0.50	18.8	0.28	29.2	0.17	47.7	0.12
22.6	0.48	19.3	0.25	33.3	0.18	45.7	0.25

23.38 Football and brain changes. American football is a contact sport that can expose players to repeated head impacts. While most impacts do not lead to a medically diagnosed concussion, an increasing body of evidence suggests that repetitive subconcussive head trauma can also result in neurocognitive deficits. Researchers recruited young football players (mean age, 12 years old). The players wore sensor-equipped helmets during all practice and game events for an entire football season to record subconcussion exposure (computed as an index value on a scale of 0 to 1). Each player also received a magnetic resonance imaging (MRI) scan at the beginning and at the end of the football season to measure changes in the brain white matter tract (computed as a percent of the original white matter volume). Here are the findings for 22 players:[14]

Exposure	Change	Exposure	Change

0.89	− 6.1	0.33	1.8
0.05	0.0	0.11	1.9
0.35	0.2	0.36	2.2
0.37	− 1.2	0.08	2.3
0.11	6.3	0.10	− 3.0
0.19	− 4.6	0.06	8.4
0.10	1.1	0.34	− 2.7
0.45	1.3	0.12	3.0
0.23	6.9	0.14	− 2.2
0.46	1.7	0.21	− 1.4
0.10	− 3.6	0.20	0.7

a. Create a scatterplot of white matter change and subconcussion exposure, using subconcussion exposure as the explanatory variable. Describe the pattern.

b. The researchers performed a linear regression analysis on the data. What is the equation for the least-squares regression line? What is the regression standard error? Test the null hypothesis of no relationship between subconcussion exposure and white matter change.

c. One player had an unusually large subconcussion exposure. Obtain the equation for the least-squares regression line without this data point and test the null hypothesis of no relationship between subconcussion exposure and white matter change. What is the regression standard error?

d. Use your work to describe the impact of the unusual observation on the regression analysis.

23.39 Testosterone and BMI in adolescent males: prediction. An adolescent male with a BMI of 30 gets a physical checkup. Using the information provided in Exercise 23.37, obtain a 95% interval to predict this adolescent’s plasma testosterone concentration.

23.40 Blood alcohol content: inference about the slope. Figure 23.8 gives the Minitab output for the regression of blood alcohol content (BAC) on number of beers consumed from Example 23.6. What is a 95% confidence interval for the slope of the regression line? Interpret this interval in context.

23.41 Blood alcohol content: inference about the intercept (optional). Figure 23.7 shows a positive linear relationship between blood alcohol content (BAC) and number of beers consumed for Example 23.6. We might expect the *intercept* α of the population regression line to be 0, because no beers consumed $(x=0)$ should produce no alcohol in the blood $(y=0)$.

a. To test

$$H_0 : \alpha = 0$$
$$H_a : \alpha \neq 0$$

we use a t statistic formed by dividing the least-squares intercept a by its standard error SE_a. Locate this statistic in the Minitab output of Figure 23.8 and verify that it is in fact a divided by its standard error. What is the reported test *P*-value? Do the data suggest that the intercept is not 0?

b. What is the 95% confidence interval for the intercept α reported by Minitab? Confidence intervals for the intercept α have the familiar form $a \pm t^* \mathrm{SE}_a$ with degrees of freedom $n-2$. Verify that the confidence interval is indeed based on this computation (for $n=17, t^*=2.13$). Explain how the confidence interval confirms your conclusion in part a.

23.42 Night watch: inference for correlation (optional). Sleeping in an unfamiliar environment can create a sleep disturbance called the first night effect. A sort of watchful sleep may occur if only one brain hemisphere truly exhibits sleep patterns. Researchers investigated asymmetric brain patterns in 11 healthy right-handed subjects on their first and second nights of sleep in the lab. The subjects' sleep latency (in minutes) was recorded each night and a noninvasive brain imaging technique was used to compute an index measuring the amount of asymmetric activity between the two brain hemispheres during deep sleep (a value of zero indicates no asymmetry). Here are the findings.[15]

Night 1	**Night 2**

Sleep latency	Asymmetry index	Sleep latency	Asymmetry index
45	− 0.107	17	0.047
12	− 0.013	5	− 0.021
26	− 0.074	9	0.001
20	0.002	1	0.004
6	0.012	19	0.006
39	− 0.055	17	0.015
30	0.019	3	0.020
22	− 0.050	5	0.027
17	− 0.045	15	− 0.037
8	− 0.009	4	0.009
9	0.001	5	0.014

a. Display the data for both nights in the same scatterplot using different symbols for each night. Alternatively, you can plot the data in two side-by-side scatterplots that use the same xx and yy scales to make both patterns truly comparable. Use sleep latency for the horizontal axis. Describe the relationship between sleep latency and brain asymmetry for each night, and then compare the patterns for both nights.

b. Taking a long time to fall asleep may indicate a greater level of sleep disturbance and explain brain asymmetry, but brain asymmetry from anxiety may manifest as increased sleep latency. Because we cannot be sure which variable truly is explanatory, obtain the correlation coefficient for each night. Use Table E to test the null hypothesis of no correlation on the first night. Do the same for the second night. What can you conclude about the brain's involvement in the first night effect?

23.43 Estimating body fat in men: inference for correlation (optional). The data file *Large.Bodyfat* contains data on the percent of body fat, age, weight, height, and ten body circumference measurements for 252 adult men. There are two assessments of a person's percent body fat (percent fat1 and percent fat2) based on two computation methods using underwater weighing. Create a scatterplot of the two variables. Neither variable is explanatory here, so

LARGE DATA SET

you should test the hypothesis of no correlation.

23.44 More on estimating body fat in men. Continue your work with the data set from the Exercise 23.43.

LARGE DATA SET

a. Create a scatterplot of percent fat1 (y)(y) as a function of abdomen circumference (x)(x) and obtain the equation for the least-squares regression.

b. Is the relationship between percent fat1 and abdomen circumference statistically significant? Run the appropriate test and check the conditions for inference.

c. A physician needs to know the average percent body fat of adult men with abdomen circumferences of 100 cm. Obtain a 95% confidence interval for your answer.

d. One adult man with abdomen circumference of 100 cm wants to predict his own percent body fat from his abdomen circumference value. Obtain a 95% prediction interval for your answer.

CHAPTER 24 One-Way Analysis of Variance: Comparing Several Means

Steve Gschmeissner/Getty Images

IN THIS CHAPTER WE COVER...

- Comparing several means
- The FF statistic
- The analysis of variance FF test
- Conditions for ANOVA
- Details of ANOVA calculations

The two-sample *t* procedures of Chapter 18 compare the means of two populations or the population mean responses to two experimental treatments. Of course, studies don't always compare just two groups. We need a method for comparing any number of means.

COMPARING SEVERAL MEANS

Comparing several populations or the population responses to several treatments starts with exploratory data analysis. When the response variable is quantitative, we use dotplots, histograms, or boxplots to describe and compare the shapes, centers, and spreads of the samples. If conditions are met, we can then use statistical inference to compare the population means. Here is an example we will explore throughout the chapter.

EXAMPLE 24.1 Rats on a cafeteria-style diet

Recent studies have suggested that palatable food high in fat, salt, and sugar can be addictive. A research team gave rats access to a cafeteria-style diet consisting of bacon, sausage, cheesecake, pound cake, frosting, and chocolate. To study the neural and behavioral impacts of extended access to such food, the researchers first randomly assigned 50 adult male Wistar rats to 3 diet groups: unlimited access to rat chow only (chow group), unlimited access to rat chow with access to cafeteria food for an hour each day (restricted group), or unlimited access to rat chow with extended access to cafeteria food for 18 hours each day (extended group). Table 24.1 gives the weights in grams (g) of the rats in each group several weeks later.[1]

TABLE 24.1 Body weights (in grams) of adult male rats fed one of three diets

Chow only									
516	547	546	564	577	570	582	594	597	599
606	606	624	623	641	655	667	690	703	
Chow plus restricted access to cafeteria food									
546	599	612	627	629	638	645	635	660	676
695	687	693	713	726	736				
Chow plus extended access to cafeteria food									
564	611	625	644	660	679	687	688	706	714
738	733	744	780	794					

STATE: We are interested in the effect of food type and access on body weight in adult male rats. Are there differences in average weight among the 3 experimental groups?

PLAN: Use graphs and numerical summaries to describe and compare the 3 distributions of body weight. Also, ask whether the differences among the mean weights of the 3 experimental groups are *statistically significant.*

SOLVE (FIRST STEPS): Figure 24.1 displays the data in dotplots. Here are the summary measures we will use in further analysis:

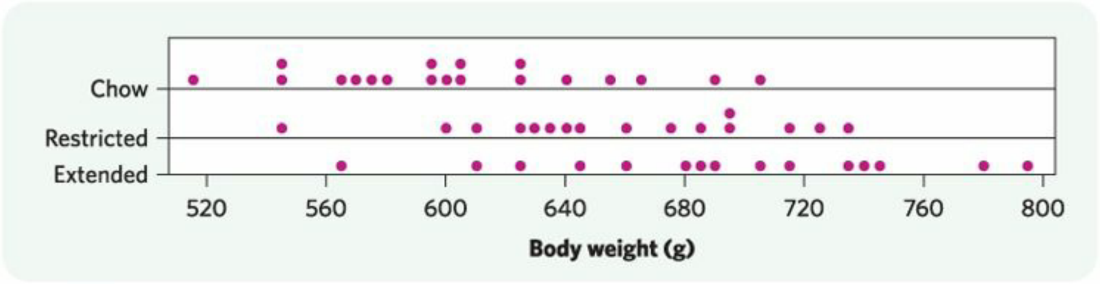


Figure 24.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 24.1 Dotplots comparing the body weights in grams of rats randomly assigned to three diet groups.

Sample	Group	Sample size	Mean weight	Standard deviation
1	Chow	19	605.6	49.64
2	Restricted	16	657.3	50.68
3	Extended	15	691.1	63.41

CONCLUDE (FIRST STEPS): The body weights in the 3 experimental groups overlap substantially. However, the weights in the extended-access group seem higher overall than the weights in the restricted-access group, which themselves seem higher overall than the weights in the chow-only group. This is reflected in mean weights of 691.1 g, 657.3 g, and 605.6 g, respectively. Are these differences in sample means statistically significant? We must develop a test for comparing more than two population means.

Call the mean weights for the 3 experimental populations of diet types μ1μ_1 for chow, μ2μ_2 for restricted, and μ3μ_3 for extended. The subscripts remind us of which group a parameter or statistic describes. The null hypothesis in statistical inference is generally a statement of “no effect.” In this study, that translates to a statement of no effect of the feeding conditions on the response variable, body weight, on average—that is, the 3 population means are equal.

STATS IN YOUR WORLD

DON’T TOUCH THE PLANTS

We know that confounding can distort inference. We don’t always recognize how easy it is to confound data. Consider the innocent scientist who visits plants in the field once a week to measure their size. A study of six plant species found that one touch a week significantly increased leaf damage by insects in two species and significantly decreased damage in another species.

In this context, we might be tempted to use the two-sample *t* test several times:

- Test H0: μ1=μ2$H_0 : \mu_1 = \mu_2$ to see if the mean weight for chow differs from the mean weight for restricted.
- Test H0: μ1=μ3$H_0 : \mu_1 = \mu_3$ to see if the mean weight for chow differs from the mean weight for extended.
- Test H0: μ2=μ3$H_0 : \mu_2 = \mu_3$ to see if the mean weight for restricted differs from the mean weight for extended.

The weakness of doing all these tests is that we get several *P*-values, one for each test performed. That doesn't tell us how likely it is that *three* sample means are spread apart as far as these values are, if the null hypothesis of no effect is true. It may be that x-1=605.6$\bar{x}_1 = 605.6$ and x-3=691.1$\bar{x}_3 = 691.1$ are significantly different if we look at just two groups, but not significantly different if we know that they are the largest and the smallest means in three groups. (Think of comparing the tallest and shortest individuals in a large group of people. They may have quite different heights, yet they simply represent two extremes of a continuum of heights.) The problem gets worse as we compare more groups, because we expect the gap between the largest and smallest sample means to get larger just by chance, even if they are all samples from the same population. Moreover, as each test comes with a chance of committing a Type I error, the chance of committing *at least one* Type I error increases with the number of tests we run. That is, *we can't safely compare many groups by doing multiple tests or confidence intervals for two parameters at a time.*

CAUTION

The problem of how to do many comparisons at once with an overall measure of confidence in all our conclusions is a common one in statistics. This is the problem of **multiple comparisons**. Statistical methods for dealing with multiple comparisons usually have two steps:

1. An *overall test* to see if there is good evidence of *any* differences among the parameters that we want to compare.
2. A detailed *follow-up analysis* to decide which of the parameters differ and to estimate how large the differences are.

The overall test, though mathematically more complex than the tests we encountered earlier, is just as straightforward when using technology. This test is called the **analysis of variance *F* test,** usually abbreviated as ANOVA. It tests the null hypothesis that there are *no differences* among the population means. The alternative hypothesis is nondirectional and simply says that H0H_0 is not true—that is, there is *some difference,* such that not all population means are equal. In other words, HaH_a states that at least one population mean differs from at least one other population mean.

EXAMPLE 24.2 Rats on a cafeteria-style diet: hypotheses

The objective of the study described in Example 24.1 is to see if, in the population of male lab rats, diet type (the explanatory variable, or factor) influences body weight (the response variable), on average. The null hypothesis is that there is no relationship between diet type and body weight—that is, there are no differences among the mean weights for the three experimental populations of diet types:

H0: μ1=μ2=μ3

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

The alternative hypothesis is that there is some difference—that is, the population means are not all the same:

Ha: Not all of μ1,μ2,and μ3are equal

$$H_a : \text{Not all of } \mu_1, \mu_2, \text{and } \mu_3 \text{are equal}$$

The alternative hypothesis is nondirectional because it allows any outcome other than “all 3 population means are equal.” For example, Ha H_a includes the case in which μ2=μ3$\mu_2 = \mu_3$ but μ1μ_1 has a different value.

When the ANOVA FF test is statistically significant, the follow-up analysis can be quite elaborate. This chapter concentrates on the overall test, along with exploratory data analysis that points to the nature of the differences. In companion Chapter 26 (available on the text’s website), we cover the details of some common follow-up analyses used when comparing several means.

APPLY YOUR KNOWLEDGE

24.1 Do fruit flies sleep? Mammals and birds sleep. Insects such as fruit flies rest, but is this rest sleep? Biologists now think that insects do sleep. One experiment gave caffeine to fruit flies to see if it affected their rest. We know that caffeine reduces sleep in mammals, so if it reduces rest in fruit flies, that’s another hint that the rest is really sleep. The paper reporting the study contains a graph similar to Figure 24.2 and states, “Flies given caffeine obtained less rest during the dark period in a dose-dependent fashion (n=36($n = 36$ per group, P<0.0001)$P < 0.0001$) .”[2]

a. The explanatory variable is amount of caffeine, in milligrams per milliliter (mg/ml) of blood. The response variable is minutes of rest (measured by an infrared motion sensor) during a 12-hour dark period. Outline the design of this experiment.

b. The *P*-value in the report comes from the ANOVA FF test. Which means does this test compare? State the null and alternative hypotheses for the test in this setting. What do the graph and the statistical test together lead you to conclude?

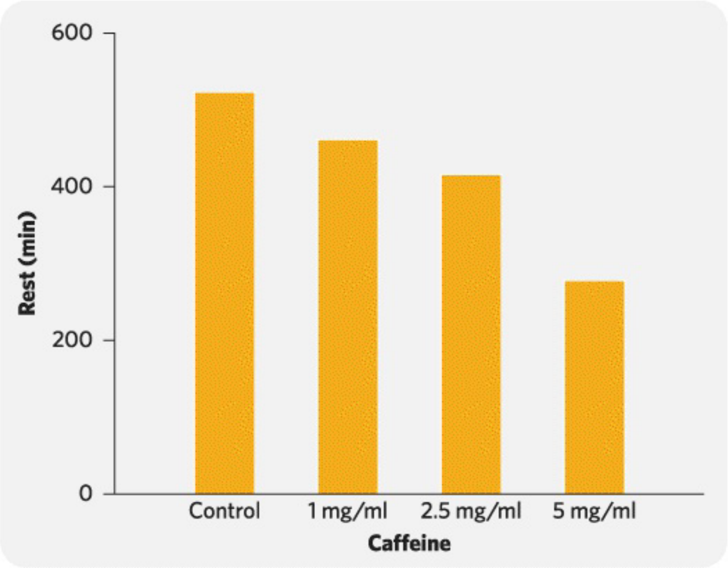


Figure 24.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 24.2 Bar graph comparing the mean rest time of fruit flies given different amounts of caffeine.

Dr. Larry Jemigan/Getty Images

24.2 Cannabis to treat brain cancer. High-grade glioma is an aggressive type of brain cancer with a low long-term survival rate. Cannabinoids, which are chemical compounds found in cannabis, are thought to inhibit glioma cell growth. Researchers transplanted glioma cells in otherwise-healthy mice, and then randomly assigned these mice to 4 cancer treatments: irradiation alone, cannabinoids alone, irradiation combined with cannabinoids, or no cancer treatment. The treatments were administered for 21 days, after which the glioma tumor volume (in cubic millimeters, mm^3) was assessed in each mouse using brain imaging. Here are the findings:[3]

Irradiation	**Cannabinoids**	**Irradiation + cannabinoids**	**Untreated**
30	12	5	24
46	14	4	30
46	16	4	43
95	41	4	51
	47	10	62
		9	32
			96

a. Name the explanatory and response variables in this experiment, and specify whether they are quantitative or categorical. Use exploratory data analysis to describe the study findings.

b. How many two-sample *t* tests would be needed to compare the treatments two at a time? Explain briefly why this approach would be problematic.

c. To compare several means, we should first perform one overall test. What are the null and alternative hypotheses for the analysis of variance FF test (ANOVA)? Be sure to specify what means the test compares.

Cinoby/Getty Images

THE *F* STATISTIC

We have stated the null and alternative hypotheses, and now we need a test statistic with a known probability distribution to help us determine whether the data differ significantly from expectations under H0.H_0. The details of ANOVA are a bit heavier computationally than those of other tests we have studied so far (they appear in an optional section at the end of this chapter). The main idea of ANOVA, however, is both more accessible and much more important.

When we have several SRSs of unrelated individuals taken under different conditions representing our explanatory variable, a null hypothesis of “no effect” of the explanatory variable on the response variable implies that the SRSs come from populations with identical distributions. Recall from Chapter 13 that sample means are less variable than individual observations, so that the sampling distribution of x-$\bar{x}$ has the N(μ,σ/n)$N(\mu,\sigma/\sqrt{n})$ distribution when sampling from a population with distribution N(μ,σ).$N(\mu,\sigma)$. Thus, even if we do not know what the values of μμ and σσ are, we know that the population distribution and the sampling distribution of x-$\bar{x}$ are related. To ask if a set of sample means gives evidence against a null hypothesis of equal population means, what matters is not just how far apart (how variable) the sample means are, but also how variable they are *relative to the variability of individual observations.*

THE IDEA OF ANALYSIS OF VARIANCE

Analysis of variance compares the variation due to specific sources (such as the levels of the explanatory variable) with the variation among individuals that should be similar (such as the individual observations in the samples).

In particular, ANOVA can test whether several populations have the same mean by comparing how much the sample means vary with how much variation there is within the samples.

It is one of the oddities of statistical language that methods for comparing means are named after the variance. The reason is that the test works by comparing two kinds of variation. Analysis of variance is a general method for studying sources of variation in responses. Comparing several means is the simplest form of ANOVA, called **one-way ANOVA**. This chapter treats only the one-way ANOVA. Companion Chapter 26 introduces the two-way ANOVA, which is used when data can be organized in two-way layouts representing two explanatory variables (two factors).

To ask whether the means of several independent SRSs differ significantly, we need a test statistic that can be used to compute a test *P*-value. In ANOVA, that test statistic is called the *F statistic*.

THE ANOVA *F* STATISTIC

The **analysis of variance** FF **statistic** for testing the equality of several means has this form:

F=variation among the sample meansvariation among individuals in the sa

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

The numerator is fundamentally a variance calculated using the sample means, while the denominator is fundamentally an average of the sample variances. Both are weighted to take into account different sample sizes. Computing the FF statistic is not mathematically challenging, but it is tedious. The computational details are available in the optional section at the end of this chapter.

The FF statistic can take only values that are zero or positive. It is zero only when all the sample means are identical and gets larger as they move farther apart. When H0H_0 is true, the numerator and the denominator should have similar values and we would expect FF to be small. FF is large when the sample means are much more variable than individuals in the same sample. Large values of FF provide evidence against the null hypothesis H0H_0 that all population means are the same. Like the alternative hypothesis Ha,H_a, the ANOVA FF test is nondirectional, because any violation of H0H_0 tends to produce a large value of F.F.

You can use the *One-Way ANOVA* applet to demonstrate the analysis of variance idea for yourself. The applet allows you to change both the group means and the spread within groups. You can watch the ANOVA FF statistic and its *P*-value change as you work.

APPLY YOUR KNOWLEDGE

24.3 ANOVA compares several means. The *One-Way ANOVA* applet displays the observations in three groups, with the group means highlighted by black dots. When you open or reset the applet, the scale at the bottom of the display shows the ANOVA FF statistic and *P*-value for these groups. (The *P*-value is marked by a red dot that moves along the scale.)

a. One group has a larger mean than the other two. Grab its mean point with the mouse. How small can you make F?F? What did you do to the mean to make FF small? Roughly how significant is your small F?F?

b. Starting with the three means aligned from your configuration at the end of part a, drag any one of the group means either up or down. What happens to F?F? What happens to the *P*-value? Convince yourself that the same thing happens if you move any one of the means, or if you move one mean slightly and then another slightly in the opposite direction.

24.4 ANOVA uses within-group variation. Reset the *One-Way ANOVA* applet to its original state.

a. Use the mouse to slide the Standard Deviation at the top of the display to the right (to increase it). You see that the group means do not change, but the spread of the observations in each group increases. What happens to FF and *P* as the spread among the observations in each group increases? What are the values of FF and *P* when you move the slider all the way to the right? Notice how variation within groups hides the differences among the group means.

b. Leave the Standard Deviation slider at the extreme right of its scale, so that spread within groups stays fixed. Use the mouse to move the group means apart. What happens to FF and P as you do this?

THE ANALYSIS OF VARIANCE *F* TEST

Hypothesis tests require a set of null and alternative hypotheses, a test statistic, and a *P*-value that gives the probability of obtaining a test statistic at least as extreme as that computed if the null hypothesis was true.

To obtain the *P*-value for an ANOVA FF test, we need to have a known probability distribution for the test statistic, F.F. Deriving this distribution mathematically is beyond the scope of an introductory statistics textbook, so we will simply state the facts. When the conditions for inference are met and the null hypothesis is true, the FF statistic follows an *F distribution.*

THE *F* DISTRIBUTIONS

The FF **distributions** are a family of distributions that take only positive values and are skewed to the right. A given FF distribution is specified by providing two distinct **degrees of freedom**.

CAUTION

The parameters of the FF distributions are the degrees of freedom of the numerator and of the denominator of the FF statistic. The numerator's degrees of freedom are always mentioned first. *Interchanging the degrees of freedom changes the distribution, so the order is important.* Our brief notation will be F(df1,df2)$F(\mathrm{df}1, \mathrm{df}2)$ for the FF distribution with df1 degrees of freedom in the numerator and df2 in the denominator. Figure 24.3 shows the density curves for three members of the FF family of distributions: F(2,30),F(5,30),$F(2, 30)$, $F(5, 30)$, and F(5,120).$F(5, 120)$.

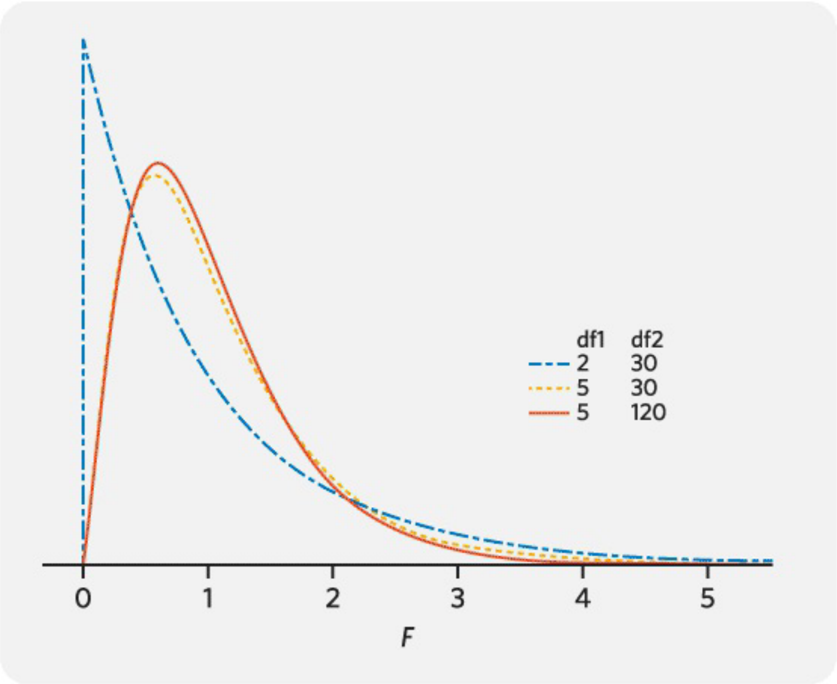


Figure 24.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 24.3 Density curves for three *F* distributions: *F*(2, 30), *F*(5, 30), and *F*(5, 120). F distributions take only positive values and are right-skewed.

Because FF distributions take only positive values, the more extreme values are always found on the right tail of the curve. When testing a null hypothesis that all population means are equal, the degrees of freedom of the ANOVA FF statistic depend on the number of means we are comparing and the number of observations in each sample.

P-VALUE OF THE ANOVA *F* TEST

We want to compare the means of kk populations. We have an SRS of size nin_i from the iith population, so that the total number of observations, NTotal,N_{Total}, in all samples combined is NTotal=n1+n2+⋯+nk.$N_{Total} = n_1 + n_2 + \cdots + n_k$.

When the null hypothesis that all population means are equal is true, the ANOVA FF statistic has the FF distribution with k-1 $k - 1$ degrees of freedom in the numerator and NTotal-k$N_{Total} - k$ degrees of freedom in the denominator.

The *P*-value of the ANOVA FF test is the area to the right of the calculated F statistic under this F distribution.

The degrees of freedom of the F statistic make sense if you think back to the definitions of the sample standard deviation and sample variance given in Chapter 2. For a sample of n data points, the variance has $n-1$ degrees of freedom. The F statistic is a ratio of two measures of variance. The numerator assesses the variance of the k sample means and so has $k-1$ degrees of freedom. The denominator assesses the variance of the data within each sample by combining information from the k sample variances (each with n_i-1 degrees of freedom) and so has $n_1-1+n_2-1+\cdots+n_k-1=N_{Total}-k$ degrees of freedom.

Software usually finds *P*-values for us. Table F in the back of the book gives critical values for several F distributions with various pairs of degrees of freedom, which can be used to find approximate *P*-values when no software is available.

EXAMPLE 24.3 Rats on a cafeteria-style diet: ANOVA

We can now complete the inference part of Example 24.1, testing the null and alternative hypotheses stated in Example 24.2.

SOLVE (INFERENCE): The software outputs in Figure 24.4 display the ANOVA results for the data of Table 24.1 from a graphing calculator, three statistical programs, and a spreadsheet program. They show that the test statistic is $F=10.71$ and the test *P*-value is $P=0.0001.$ There is very strong, highly statistically significant evidence that the three experimental conditions do not all lead to the same population mean weight.

The F test does not say *which* of the three population means are significantly different. Our preliminary data analysis indicates that, on average, rats in the extended-access group had the highest weights and that rats in the chow-only group had the lowest weights at the end of this study.

CONCLUDE: There is very strong evidence $(P=0.0001)$ that

> the imposed feeding conditions affect body weight, on average, in the population of adult male Wistar rats. The most important difference recorded among the sample means is that rats given extended access to cafeteria food ended up being heavier than rats given access only to chow, on average.

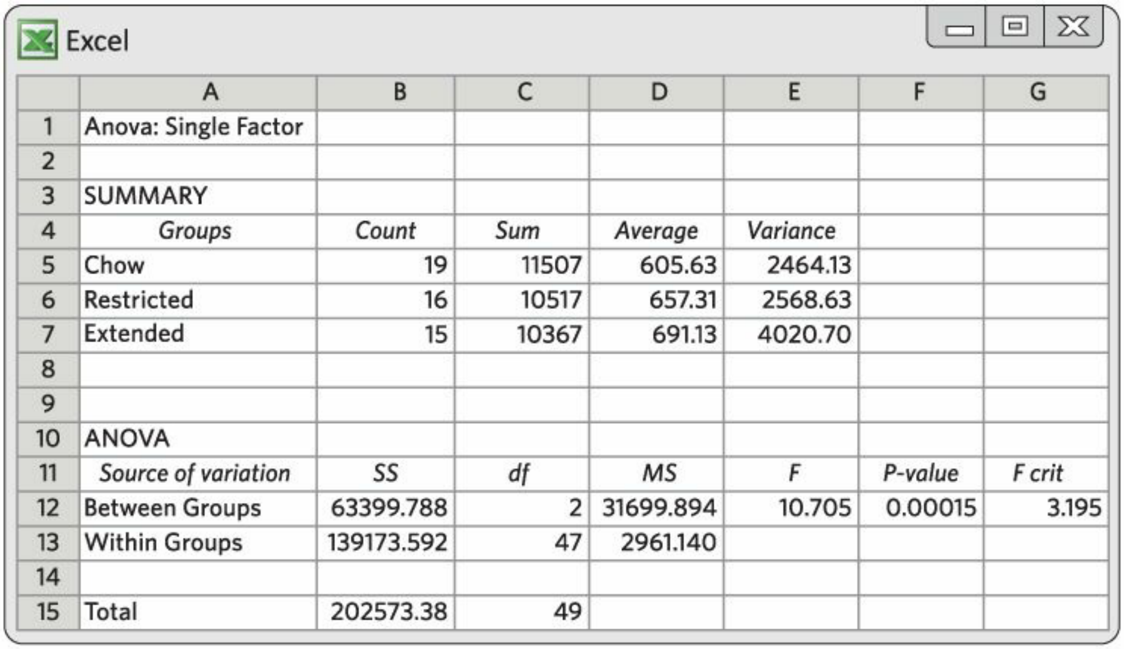
The computations performed by software in this example involve the following steps:

- Summary statistics: x-,s,$\bar{x}, s,$ and nn for each sample (displayed in Example 24.1)
- Degrees of freedom of the FF distribution: The three sample sizes are n1=19,n2=16,n3=15,$n_1 = 19, n_2 = 16, n_3 = 15,$ so the total number of observations is NTotal=50.$N_{Total} = 50.$ The ANOVA FF statistic has numerator degrees of freedom k-1=3-1=2,$k - 1 = 3 - 1 = 2,$ and denominator degrees of freedom NTotal-k=50-3=47. $N_{Total} - k = 50 - 3 = 47.$
- FF statistic: F=10.71$F = 10.71$ (The FF statistic is computed using the summary statistics and degrees of freedom, as shown in the optional section at the back of the book.)
- *P*-value: This is the area to the right of F=10.71,$F = 10.71,$ under the F(2,47)$F\,(2, 47)$ distribution. The TI-83 calculator command tcdf(10.71,1E99,2,47) and the R command pf(10.71, df1=2, df2=47, lower.tail=FALSE) both confirm that the *P*-value is 0.000147.

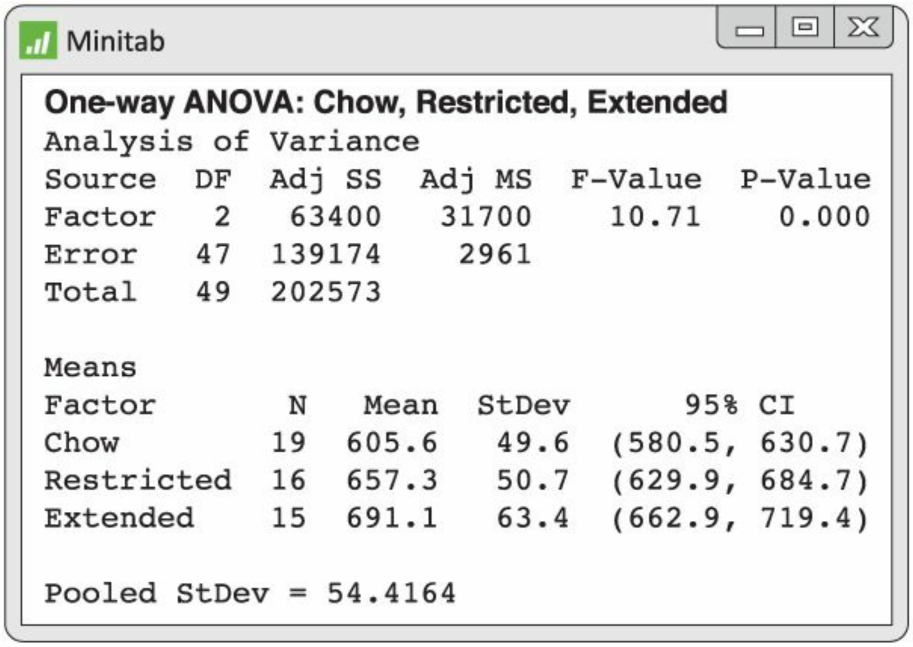
  Table F at the back of the book does not have critical values for the F(2,47)$F\,(2, 47)$ distribution, but the F(2,50)$F\,(2, 50)$ distribution is similar enough to be used as an approximation. Because the FF statistic is greater than the largest critical value provided for this distribution, we conclude from Table F that the *P*-value is smaller than 0.001.

pp	.001
F*F^*	7.96

The technology outputs in Figure 24.4 are quite varied. All give the FF test statistic and the *P*-value, rounded to a varying number of decimal places. All but the graphing calculator provide summary statistics. Minitab, JMP, and SPSS also give confidence intervals for each mean separately. (We will see on page 624 how these confidence intervals are derived.) The Chow and Restricted intervals overlap only slightly, and the Extended interval is well above that for Chow. However, beware that we are *not* 95% confident that *all three* intervals capture their respective population means. A different approach is required for multiple comparisons so that we can be 95% confident that *all* our conclusions are correct. Companion Chapter 26 shows you how to perform and interpret such follow-up analysis.



	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Chow	19	11507	605.63	2464.13		
6	Restricted	16	10517	657.31	2568.63		
7	Extended	15	10367	691.13	4020.70		
8							
9							
10	ANOVA						
11	*Source of variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between Groups	63399.788	2	31699.894	10.705	0.00015	3.195
13	Within Groups	139173.592	47	2961.140			
14							
15	Total	202573.38	49				

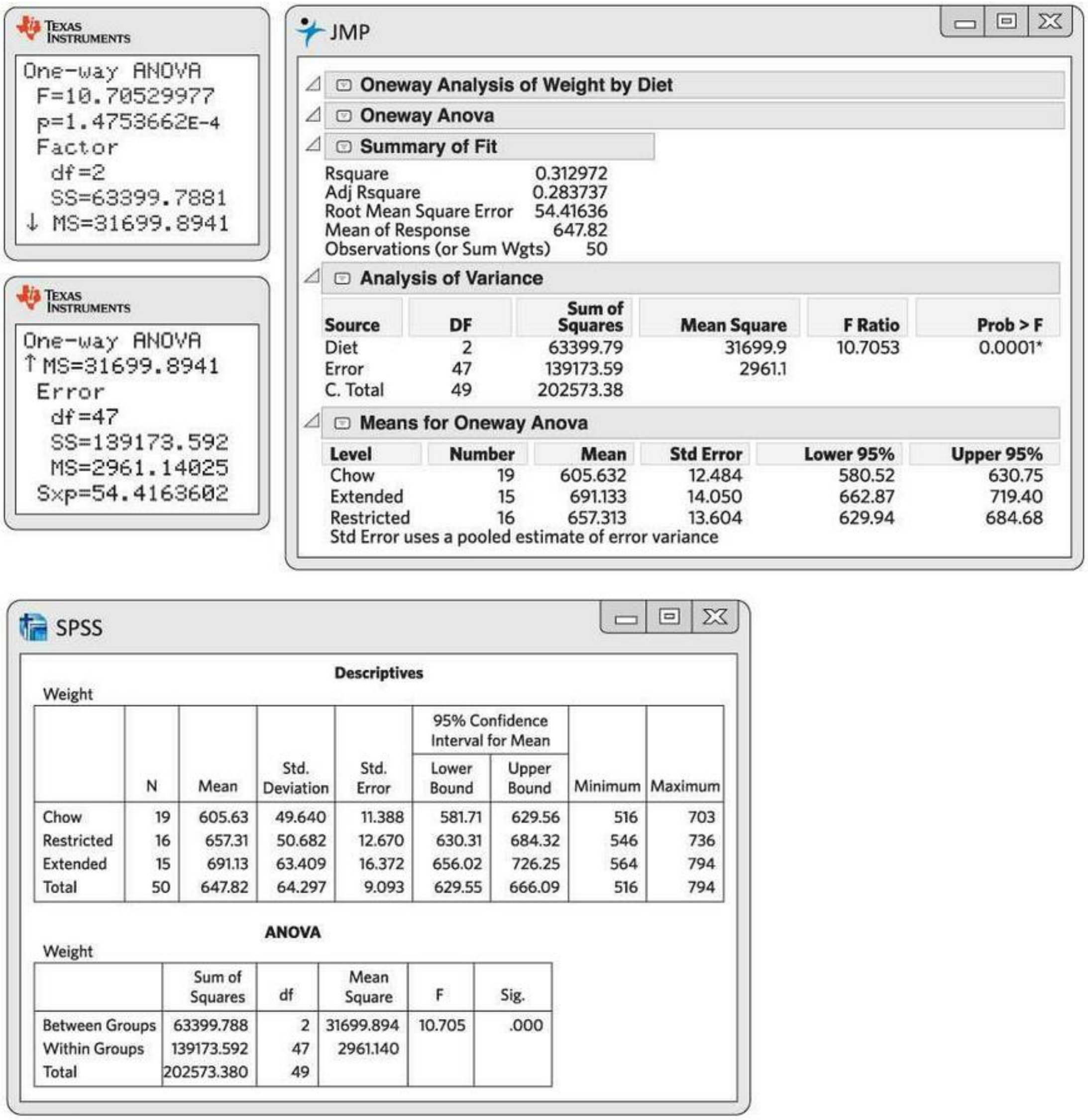


One-way ANOVA: Chow, Restricted, Extended

```
Analysis of Variance
Source  DF  Adj SS  Adj MS  F-Value  P-Value
Factor   2   63400   31700    10.71    0.000
Error   47  139174    2961
Total   49  202573

Means
Factor       N   Mean  StDev       95% CI
Chow        19  605.6   49.6  (580.5, 630.7)
Restricted  16  657.3   50.7  (629.9, 684.7)
Extended    15  691.1   63.4  (662.9, 719.4)

Pooled StDev = 54.4164
```

Figure 24.4a

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



```
One-way ANOVA
 F=10.70529977
 p=1.4753662E-4
 Factor
  df=2
  SS=63399.7881
↓ MS=31699.8941
```

```
One-way ANOVA
↑MS=31699.8941
 Error
  df=47
  SS=139173.592
  MS=2961.14025
 Sxp=54.4163602
```

Oneway Analysis of Weight by Diet

Oneway Anova

Summary of Fit

Rsquare	0.312972
Adj Rsquare	0.283737
Root Mean Square Error	54.41636
Mean of Response	647.82
Observations (or Sum Wgts)	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Diet	2	63399.79	31699.9	10.7053	0.0001*
Error	47	139173.59	2961.1		
C. Total	49	202573.38			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Chow	19	605.632	12.484	580.52	630.75
Extended	15	691.133	14.050	662.87	719.40
Restricted	16	657.313	13.604	629.94	684.68

Std Error uses a pooled estimate of error variance

Descriptives

Weight

					95% Confidence Interval for Mean			
	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
Chow	19	605.63	49.640	11.388	581.71	629.56	516	703
Restricted	16	657.31	50.682	12.670	630.31	684.32	546	736
Extended	15	691.13	63.409	16.372	656.02	726.25	564	794
Total	50	647.82	64.297	9.093	629.55	666.09	516	794

ANOVA

Weight

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	63399.788	2	31699.894	10.705	.000
Within Groups	139173.592	47	2961.140		
Total	202573.380	49			

Figure 24.4b

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 24.4 ANOVA for the rat weight data: output from a graphing calculator, three statistical programs, and a spreadsheet program.

APPLY YOUR KNOWLEDGE

24.5 Logging in the rainforest. How does logging in a tropical rainforest affect the forest several years later? Researchers compared forest plots in Borneo that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). Although the study was not an experiment, the authors explain why we can consider the plots to be randomly selected. The data appear in Table 24.2. The variable Trees is the count of trees in a plot; Species is the count of tree species in a plot. The variable Richness is the number of species divided by the number of individual trees, Species/Trees.[4]

TABLE 24.2 **Data from a study of logging in Borneo**

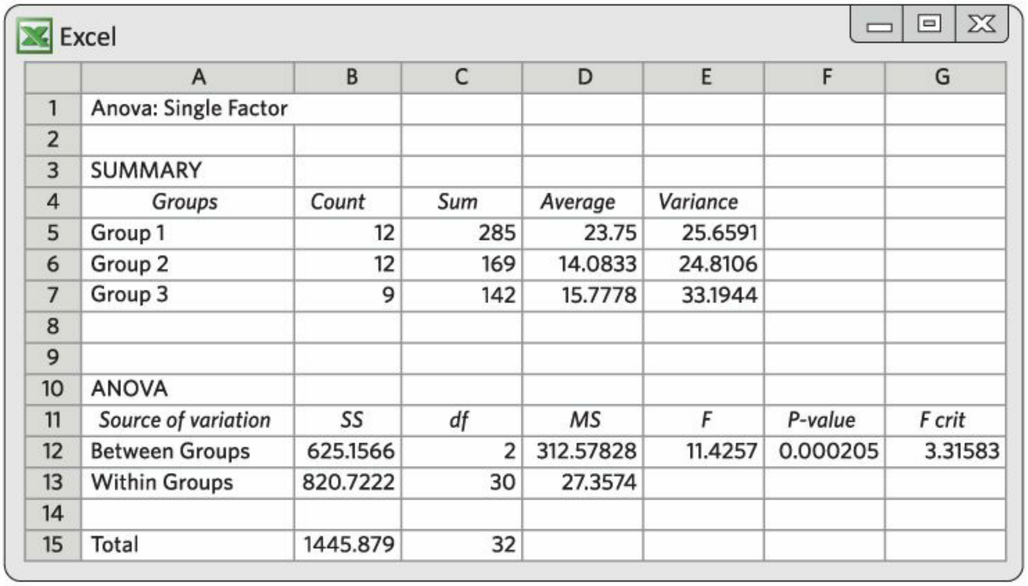
Observation	Group	Trees	Species
1	1	27	22
2	1	22	18
3	1	29	22
4	1	21	20
5	1	19	15
6	1	33	21
7	1	16	13
8	1	20	13
9	1	24	19
10	1	27	13
11	1	28	19
12	1	19	15
13	2	12	11
14	2	12	11
15	2	15	14
16	2	9	7
17	2	20	18
18	2	18	15
19	2	17	15
20	2	14	12
21	2	14	13
22	2	2	2
23	2	17	15
24	2	19	8
25	3	18	17
26	3	4	4
27	3	22	18
28	3	15	14

29	3	18	18
30	3	19	15
31	3	22	15
32	3	12	10
33	3	12	12

a. Display the Trees data for the three groups in dotplots sharing the same scale. What effects of logging are noticeable?

b. Figure 24.5 shows Excel ANOVA output for Trees. What do the group means show about the effects of logging?

c. What are the values of the ANOVA FF statistic and its *P*-value? What hypotheses does FF test? What conclusions about the effects of logging on number of trees do the data lead to?



	A	B	C	D	E	F	G
1	Anova: Single Factor						
2							
3	SUMMARY						
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5	Group 1	12	285	23.75	25.6591		
6	Group 2	12	169	14.0833	24.8106		
7	Group 3	9	142	15.7778	33.1944		
8							
9							
10	ANOVA						
11	*Source of variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Between Groups	625.1566	2	312.57828	11.4257	0.000205	3.31583
13	Within Groups	820.7222	30	27.3574			
14							
15	Total	1445.879	32				

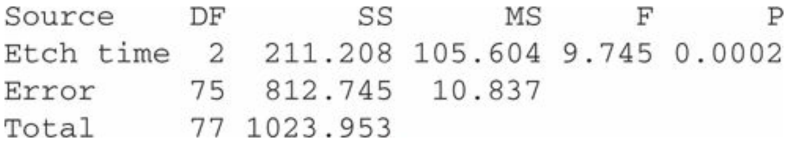
Figure 24.5

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 24.5 Excel ANOVA output for the number of trees in forest plots in Borneo.

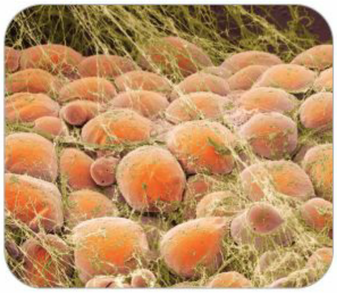
24.6 Adipocyte size. Obesity is characterized by an abnormal increase in the number and size of adipocytes, fat-storing cells that enlarge to accommodate surplus energy. Muscadine grape phytochemical (MGP) is a dietary supplement thought to have some health benefits, so researchers explored its potential in fighting obesity. They randomly assigned male mice to 3 diets: low-fat (LF), high-fat (HF), or a high-fat diet supplemented with MGP (HF+MGP). After 15 weeks, the typical adipocyte diameter was determined by microscope evaluation for each mouse. Here are the findings (in microns), followed by the ANOVA output from the statistical software R:[5]

LF	85	84	95	87	101
HF	108	100	93	85	107
HF+MGP	83	88	91	79	84



```
Source      DF        SS      MS     F      P
Etch time    2   211.208 105.604 9.745 0.0002
Error       75   812.745  10.837
Total       77  1023.953
```

a. Plot the data and obtain the summary statistics. What does your exploratory data analysis suggest about the effect of diet type on adipocyte size?
b. What are the null and alternative hypotheses for this ANOVA test? What are the values of the FF statistic and the *P*-value? If you have technology capable of performing an ANOVA FF test, verify that you obtain the same values.
c. Assuming that the conditions for ANOVA are met, conclude in context.



Steve Gschmeissner/Getty Images

CONDITIONS FOR ANOVA

Like all inference procedures, ANOVA is valid only in some circumstances. Here are the conditions under which we can use ANOVA to compare population means.

CONDITIONS FOR APPLYING ANOVA

- We have k **independent SRSs,** one from each of k populations.
- Each of the k populations has a **Normal distribution** with an unknown mean. The parameter μ_i is the unknown mean of the ith population. The means may be different in the different populations.
- All the populations have the **same standard deviation** σ, whose value is unknown.

The first two requirements are familiar from our study of the two-sample *t* procedures for comparing two means. As usual, the design of the data production is the most important condition for inference. Biased sampling or confounding can make any inference meaningless. *If we do not actually draw separate SRSs from each population or carry out a randomized comparative experiment, it may be unclear to which population the conclusions of inference apply.* In the life sciences, samples are often not true SRSs, but rather individuals randomly chosen from a pool of available individuals. You must judge each use on its merits—a judgment that usually requires some knowledge of the subject of the study in addition to some knowledge of statistics.

Because no real population has an exactly Normal distribution, the usefulness of inference procedures that assume Normality depends on how sensitive they are to departures from Normality. Fortunately, procedures for comparing means are not very sensitive to lack of Normality. The ANOVA F test, like the *t* procedures, is **robust**. What matters is Normality of the sample means, so ANOVA becomes safer as the sample sizes get larger, because of the central limit theorem effect. Remember to check for outliers that change the value of sample means and for extreme skewness. When there are no outliers and the distributions are roughly symmetric, you can safely use ANOVA for sample sizes as small as 4 or 5 for each group.

The third condition is more of a challenge: ANOVA assumes that the variability of observations, measured by the standard deviation, is the same in all populations. As discussed in Chapter 18 (page 462), there is a special version of the two-sample *t* test that assumes equal standard deviations in both populations. The ANOVA FF statistic for comparing two means is exactly the square of this special *t* statistic. We prefer the *t* test that does not assume equal standard deviations, but for comparing more than two means there is no general alternative to the ANOVA F.F. It is not easy to check the condition that the populations have equal standard deviations. Statistical tests for equality of standard deviations are very sensitive to lack of Normality, so much so that they are of little practical value. You must either seek expert advice or rely on the robustness of ANOVA.

How serious are the effects of unequal standard deviations? ANOVA is not too sensitive to violations of the condition, especially when all samples have the same or similar sizes and no sample is very small. When designing a study, try to take samples of roughly the same size from all the groups you want to compare. The sample standard deviations estimate the population standard deviations, so check before doing ANOVA that the sample standard deviations are similar to each other. We expect some variation among them due to chance. Here is a conservative rule of thumb that is safe in almost all situations.

CHECKING STANDARD DEVIATIONS IN ANOVA

The results of the ANOVA FF test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation.

EXAMPLE 24.4 Rats on a cafeteria-style diet: conditions for ANOVA

The rat diet study is based on the random assignment of adult male

Wistar lab rats to three treatment groups, which creates independent random samples that the researchers consider to be representative of all adult male Wistar rats. The dotplots in Figure 24.1 show no obvious departure from Normality for the weights in each group. The sample standard deviations for the three experimental groups are

$$s1=49.64 \quad s2=50.68 \quad s3=63.41$$

$$s_1 = 49.64 \quad s_2 = 50.68 \quad s_3 = 63.41$$

These standard deviations satisfy our conservative guideline:

$$\text{largest ssmallest } s=63.4149.64=1.28$$

$$\frac{\text{largest } s}{\text{smallest } s} = \frac{63.41}{49.64} = 1.28$$

We can safely use ANOVA to compare the mean weights for the three experimental populations.

EXAMPLE 24.5 Smoking during pregnancy

STATE: Cigarette labels warn pregnant women against smoking. Does nicotine actually reach the fetus, crossing the protective placental barrier? Researchers selected consecutive pregnant women delivering at an Egyptian hospital and categorized them as active smokers, passive smokers, or nonsmokers. They then analyzed the newborns’ meconium for cotinine content, the metabolized form of nicotine. Meconium is a newborn’s first stool right after birth, so it is a good indicator of fetal exposure to drugs or other chemical agents. Here are the meconium cotinine levels (in nanograms per milliliter, ng/ml) for the three groups:[6]

Active smokers	490	418	405	328	700	292	295	272	240	232
Passive smokers	254	219	287	257	271	282	148	273	350	293
Nonsmokers	158	163	153	207	211	159	199	187	200	213

PLAN: Plot and summarize the data to assess the effect of smoking exposure on meconium cotinine levels, and check that we can safely use ANOVA to test the hypotheses

H0: μ1=μ2=μ3Ha: Not all of μ1,μ2,and μ3 are equal

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_a : \text{Not all of } \mu_1, \mu_2, \text{and } \mu_3 \text{ are equal}$$

where μ1,μ2,$\mu_1, \mu_2,$ and μ3μ_3 are the mean meconium cotinine levels in the populations of infants born to mothers who are, respectively, active smokers, passive smokers, and nonsmokers.

SOLVE: EXPLORATORY DATA ANALYSIS. Figure 24.6 displays the data in dotplots with a common axis. Here are the summary statistics:

Sample	Group	Sample size	Mean	Standard deviation
1	Active smoker	10	367.2	143.7
2	Passive smoker	10	263.4	52.5
3	Nonsmoker	10	185.0	24.2

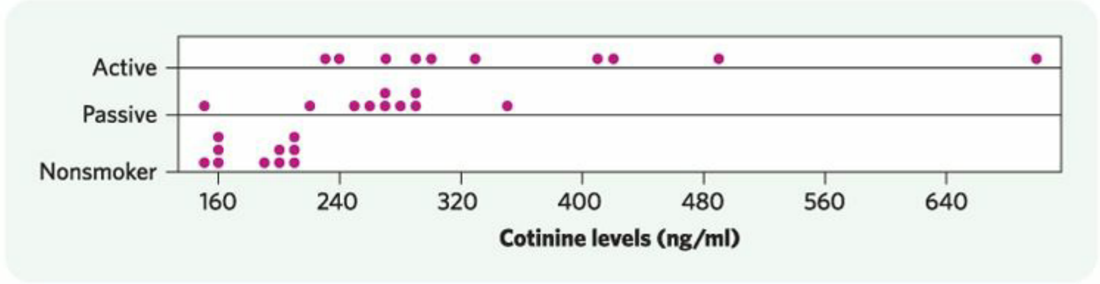


Figure 24.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 24.6 Dotplots comparing the meconium cotinine levels of infants born to mothers with different smoking statuses.

The graph shows clearly that the infants born to nonsmoking mothers had overall much lower meconium levels of cotinine, suggesting that a mother's nicotine exposure may indeed reach the fetus. In fact, the mean cotinine level was approximately twice as high for those infants in the study who were born to active smokers than for those born to nonsmokers. To test whether the mean meconium cotinine level is not the same in all three populations, we need to check that the conditions for inference are met.

SOLVE: CHECK THE CONDITIONS FOR INFERENCE.

- We don't have actual SRSs from the populations of interest. However, because the selection method appears to be unbiased and the metabolic processing of nicotine is likely similar for all pregnant women, we may be willing to consider these women as representative of the larger population of pregnant women. The women were then divided by smoking status, so we have 3 independent random samples, suitable for ANOVA.

- Figure 24.6 shows that the active smokers group is clearly skewed with an outlier. The other groups appear to be roughly symmetric. ANOVA assumes Normally distributed populations but is robust to deviations from that assumption for large enough sample sizes. With a total sample size of $N_{Total} = 30$ equally divided across the 3 random samples, the noted departures from Normality should not be a problem.
- ANOVA requires sample standard deviations that are similar enough to meet the assumption of equal population standard deviation. The ratio

$$\frac{\text{largest } s}{\text{smallest } s} = \frac{143.7}{24.2} = 5.9$$

is much more than 2, so these data clearly do not satisfy our rule of thumb.

CONCLUDE: The sample data suggest that mean meconium cotinine levels are not the same regardless of the mother’s smoking status. However, an ANOVA F test is not appropriate for these data because of the pronounced disparities in standard deviations.

It should come as no surprise that infants born to mothers with no exposure to nicotine have consistently low cotinine levels, and that cotinine levels are much more variable among infants born to mothers who actively smoke (because the mothers’ smoking amounts probably vary quite a lot as well). In their published report, the researchers wrote that they ran an ANOVA and reported that $F = 10.45$ and $P < 0.01$. Those are indeed the values that any statistical software provide if you use ANOVA on the data, yet they are meaningless because the requirements for ANOVA are not met. *It is one of the drawbacks of technology that it will perform statistical computations regardless of their validity. It is up to you to carefully examine your data to check that the analysis you selected is appropriate for your data and that the conditions for inference are met.*

In fact, there is more than one inference procedure capable of comparing the distributions of a quantitative response variable in several populations. This chapter introduced the ANOVA FF test, a very commonly used inference procedure for such situations. Other procedures approach the problem from a different perspective, which does not require Normality or equal population standard deviation. Chapter 27 on the companion website describes one such method (the Kruskal-Wallis test) that uses the ranks of the data points rather than the individual values.

APPLY YOUR KNOWLEDGE

24.7 Logging in the rainforest, continued. Refer to your exploratory data analysis for Exercise 24.5 and the software output in Figure 24.5 to determine whether ANOVA is appropriate to compare the mean number of trees in these three populations.

24.8 Adipocyte size, continued. Refer to your exploratory data analysis for Exercise 24.6 to determine whether ANOVA is appropriate to test whether the mean adipocyte diameter of male mice is affected by the experimental diets.

24.9 Cannabis to treat brain cancer, continued. In Exercise 24.2 you performed the exploratory data analysis for an experimental study on the effect of four cancer treatments on glioma tumor volume. You also stated the null and alternative hypotheses for an ANOVA FF test. Explain why the conditions for ANOVA are not met.

DETAILS OF ANOVA CALCULATIONS[i]

We typically use software to perform the computations of inference procedures. This is especially true for ANOVA, because of its lengthy computations. However, looking at formulas can help us understand the basis for the computations. Also, there may be times when using statistical software on the raw data is not an option. This section provides the general formula for the ANOVA FF statistic that we have skipped so far.

We have SRSs from each of kk populations. Subscripts from 1 to kk tell us which sample a statistic refers to:

Population	Sample size	Sample mean	Sample std. dev.
1	n1n_1	x-1$\bar{x}_1$	s1s_1
2	n2n_2	x-2$\bar{x}_2$	s2s_2
⋮	⋮	⋮	⋮
kk	nkn_k	x-k$\bar{x}_k$	sks_k

You can find the FF statistic from just the sample sizes ni,n_i, the sample means x-i,$\bar{x}_i$, and the sample standard deviations si.s_i. You don't need to go back to the individual observations.

The ANOVA FF statistic has the form

F=variation among the sample meansvariation among individuals in the same sample

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

The measures of variation in the numerator and denominator of FF are called **mean squares**. A mean square is a more general form of a sample variance. An ordinary sample variance s2s^2 is an average (or mean) of the squared deviations of observations from their mean, so it qualifies as a "mean square."

The numerator of FF is a mean square that measures variation among the kk sample means x-1,x-2,…,x-k.$\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$. Call x-$\bar{x}$ the overall mean response (the mean of all NTotalN_{Total} observations together). You can find x-$\bar{x}$ from the kk sample means by taking x-=n1x-1+n2x-2+···+nkx-kNTotal

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{N_{Total}}$$

The sum of each mean multiplied by the number of observations it represents is the sum of all the individual observations. Dividing this sum by NTotal,N_{Total}, the total number of observations in the study, gives the overall mean x-.$\bar{x}$. The numerator mean square in FF is an average of the kk squared deviations of the means of the samples from the overall x-.$\bar{x}$. We call it the **mean square for groups**, abbreviated as **MSG**:

MSG=n1(x-1-x-)2+n2(x-2-x-)2+···+nk(x-k-x-)2k-1

$$\text{MSG} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2}{k-1}$$

Each squared deviation is weighted by ni,n_i, the number of observations it represents.

The mean square in the denominator of FF measures variation among individual observations in the same sample. For any one sample, the sample variance si2s_i^2 does this job. For all kk samples together, we use an average of the individual sample variances. It is again a weighted average in which each si2s_i^2 is weighted by one fewer than the number of observations it represents, ni-1.$n_i - 1$. Put another way, each si2s_i^2 is weighted by its degrees of freedom, ni-1.$n_i - 1$. The resulting mean square is called the **mean square for error, MSE**:

MSE=(n1-1)s12+(n2-1)s22+···+(nk-1)sk2NTotal-k

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2}{N_{Total} - k}$$

“Error” doesn’t mean that a mistake has been made. It’s a traditional term for chance variation. Here is a summary of the ANOVA test.

FORMULA OF THE ANOVA *F* STATISIC

Draw an independent SRS from each of kk populations. The ii th sample has size ni,n_i, sample mean x-i,$\bar{x}_i$, and sample standard deviation si.s_i.

The **ANOVA** FF **statistic** for comparing kk population means is

F=MSGMSE

$$F = \frac{\text{MSG}}{\text{MSE}}$$

The numerator of FF is the **mean square for groups**:

MSG=n1(x-1-x-)2+n2(x-2-x-)2+···+nk(x-k-x-)2k-1

$$\text{MSG} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2 + \cdots + n_k(\bar{x}_k - \bar{x})^2}{k-1}$$

The denominator of FF is the **mean square for error**:

MSE=(n1-1)s12+(n2-1)s22+···+(nk-1)sk2NTotal-k

$$\text{MSE} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \cdots + (n_k-1)s_k^2}{N_{Total} - k}$$

The denominators in the formulas for MSG and MSE are the two degrees of freedom, k-1$k-1$ and NTotal-k,$N_{Total} - k$, of the FF test. The mean square numerators are called **sums of squares**, from their algebraic form. The usual practice is to present the results of ANOVA in an **ANOVA table**. Output from software usually includes an ANOVA table.

EXAMPLE 24.6 Rats on a cafeteria-style diet: ANOVA table

Look again at the five outputs in Figure 24.4. The four software outputs give the ANOVA table. The TI-83, with its small screen, gives the same information as a list rather than a table. Each output uses slightly different language to identify the two sources of variation. The basic ANOVA table is

Source of variation	**df**	**SS**	**MS**	***F* statistic**
Among samples (groups)	2	63,399.79	MSG=31,699.89$\text{MSG} = 31{,}699.89$	10.705
Within samples	47	139,173.59	MSE=2,961.14$\text{MSE} = 2{,}961.14$	

> (error)
>
> You can check that each mean square MS is the corresponding sum of squares SS divided by its degrees of freedom df. The FF statistic is MSG divided by MSE.

The step-by-step computations, using the summary statistics provided in Example 24.1, are as follows:

- Overall mean of the 50 rat weights in the study:

x-=n1x-1+n2x-2+n3x-3NTotal=(19)(605.63)+(16)(657.31)+(15)(691.13)50=32, 390.8850=647.82

$$\begin{aligned} \bar{x} &= \frac{n_1\bar{x}_1+n_2\bar{x}_2+n_3\bar{x}_3}{N_{Total}} \\ &= \frac{(19)(605.63)+(16)(657.31)+(15)(691.13)}{50} \\ &= \frac{32,390.88}{50}=647.82 \end{aligned}$$

- Mean square for groups:

MSG=n1(x-1-x-)2+n2(x-2-x-)2+n3(x-3-x-)2k-1=13-1[(19)(605.63-647.82)2+(16)(657.31-647.82)2+(15)(691.13-647.82)2]=63, 397.232=31,698.62

$$\begin{aligned} \text{MSG} &= \frac{n_1(\bar{x}_1-\bar{x})^2+n_2(\bar{x}_2-\bar{x})^2+n_3(\bar{x}_3-\bar{x})^2}{k-1} \\ &= \frac{1}{3-1}\Big[(19)\,(605.63-647.82)^2+(16)\,(657.31-647.82)^2 \\ &\qquad +(15)\,(691.13-647.82)^2\Big] \\ &= \frac{63,397.23}{2}=31,698.62 \end{aligned}$$

- Mean square for error:

MSE=(n1-1)s12+(n2-1)s22+(n3-1)s32NTotal-k=(18)(49.642)+(15)(50.682)+(14)(63.412)47=139, 172.8647=2,961.12

$$\begin{aligned}\text{MSE} &= \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+(n_3-1)s_3^2}{N_{Total}-k}\\ &= \frac{(18)(49.64^2)+(15)(50.68^2)+(14)(63.41^2)}{47}\\ &= \frac{139,172.86}{47}=2,961.12\end{aligned}$$

- FF statistic:

F=MSGMSE=31, 698.622, 961.12=10.705

$$F=\frac{\text{MSG}}{\text{MSE}}=\frac{31,698.62}{2,961.12}=10.705$$

These computations differ slightly in places from the output in Figure 24.4 because of roundoff error.

Because MSE is an average of the individual sample variances, it is also called the *pooled sample variance*, written as sp2.s_p^2. When all kk populations have the same population variance σ2,σ^2, as ANOVA assumes that they do, sp2s_p^2 estimates the common variance σ2.σ^2. The square root of MSE (root mean square error) is the **pooled standard deviation** sp.s_p. It estimates the common standard deviation σσ of observations in each group. The Minitab, JMP, and TI-83 outputs in Figure 24.4 give the value sp=54.416.$s_p=54.416$.

The pooled standard deviation sps_p is a better estimator of the common σσ than any individual sample standard deviation sis_i because it combines (pools) the information in all kk samples. We can get a confidence interval for any of the means μiμ_i from the usual form

estimate±t*SEestimate

$$\text{estimate} \pm t^*\text{SE}_{\text{estimate}}$$

using sps_p to estimate σ.σ. The confidence interval for μiμ_i is

x-i±t*spni

$$\bar{x}_i \pm t^*\frac{s_p}{\sqrt{n_i}}$$

Use the critical value t*t^* from the *t* distribution with NTotal-k$N_{Total}-k$ degrees of freedom, because sps_p has NTotal-k$N_{Total}-k$ degrees of freedom. These are the confidence intervals that appear in the Minitab and JMP outputs in Figure 24.4 (whereas SPSS provides regular one-sample *t* confidence intervals).

Lastly, the one-way ANOVA tests the hypothesis that all of the k populations examined have the same mean. When there are just two populations, the hypotheses become

H0: μ1=μ2Ha: μ1≠μ2

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_a &: \mu_1 \neq \mu_2 \end{aligned}$$

which can also be tested with a two-sample *t* procedure.

Because ANOVA requires equal population standard deviations, it is most closely related to the pooled two-sample *t* test. Recall from Chapter 18 (page 462) that the pooled two-sample *t* statistic is

t=x-1-x-2sp1n1+1n2

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s_p, the pooled estimate of the population standard deviation σ, is

sp=(n1-1)s12+(n2-1)s22n1+n2-2

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

In fact, in the two-sample case the F statistic is exactly the square of the pooled *t* statistic:

F=variation among the sample meansvariation among individuals within sa
(x¯1−x¯2)2sp2(1n1+1n2)=t2

$$\begin{aligned} F &= \frac{\text{variation among the sample means}}{\text{variation among individuals within samples}} \\ &= \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = t^2 \end{aligned}$$

Both tests give exactly the same two-sided *P*-value.

In Chapter 18 we discussed the advantages of not having to assume equal population variances. When comparing two means, we recommend avoiding the pooled *t* procedure (or, equivalently, ANOVA) because extensive studies show that *t* with its accurate degrees of freedom performs essentially as well as pooled *t* when the standard deviations really are equal and provides notably more accurate *P*-values when they are not. When we want to compare more than two means, however, there is no simple way to avoid the equal standard deviations condition.

APPLY YOUR KNOWLEDGE

The calculations of ANOVA use only the sample sizes n_i*, the sample means* $\bar{x}_i$*, and the sample standard deviations* s_i*. You can therefore recreate the ANOVA calculations when a report gives these summaries but does not provide the actual data. These optional exercises ask you to do the ANOVA calculations starting with the summary statistics.*

Scott Camazine/Science Source

24.10 Exercise type and body fat. Obesity and overweight are caused by excessive adipose tissue, or body fat. Visceral fat, surrounding the internal organs, is clearly associated with heart disease and diabetes. Subcutaneous fat, found just below the skin (often in the buttocks and thighs), is not. A study examined the impact of exercise type on visceral and subcutaneous fat. Overweight, sedentary, but otherwise disease-free adults were randomly assigned to 3 exercise regimens for eight months: aerobic training, resistance training, aerobic plus resistance training. All exercise sessions were supervised to ensure correct completion. The subjects' body fat amount was assessed with computed tomography imaging (in cm^2) at the beginning and at the end of the experiment. The study report contains the following information about the visceral fat reduction (in cm^2) achieved by the subjects in each group (note that the reduction is indicated as a positive value):[7]

Treatment	**n**	**$\bar{x}$**	**s**
Aerobic training	36	15.9	34

Resistance training	39	0.8	19
Aerobic plus resistance training	35	10.9	33

a. Do the standard deviations satisfy the rule of thumb for safe use of ANOVA?

b. The report does not provide the distributions of visceral fat reduction. ANOVA is nonetheless safe for these data. Why?

c. Calculate the overall mean response x-,$\bar{x}$, the mean square for groups (MSG), and the mean square for error (MSE).

d. Obtain the ANOVA FF statistic and the test *P*-value. Is there evidence that the mean visceral fat reduction in overweight adults depends on which 3 exercise programs they follow?

24.11 Exercise type and body fat, continued. The study in Exercise 24.10 also reports the subcutaneous fat reduction (in cm^2) achieved by the subjects in each group (note that the reduction is indicated as a positive value):

Treatment	**n**n	**x-**$\bar{x}$	**s**s
Aerobic training	36	25.1	30
Resistance training	39	8.2	54
Aerobic plus resistance training	35	28.7	37

a. Do the standard deviations satisfy the rule of thumb for safe use of ANOVA?

b. The report does not provide the distributions of subcutaneous fat reduction. ANOVA is nonetheless safe for these data. Why?

c. Calculate the overall mean response x-,$\bar{x}$, the mean square for groups (MSG), and the mean square for error (MSE).

d. Obtain the ANOVA FF statistic and the test *P*-value. Is there evidence that the mean subcutaneous fat reductions in overweight adults depends on which 3 exercise programs they follow?

Chapter 24 SUMMARY

- **One-way analysis of variance (ANOVA)** compares the means of several populations. The **ANOVA** FF **test** tests the overall H0H_0 that all the populations have the same mean. If the FF test shows significant differences, examine the data to see where the differences lie and whether they are large enough to be important.
- The **conditions for ANOVA** state that we have an **independent SRS** from each population; that each population has a **Normal distribution**; and that all populations have the **same standard deviation**.
- In practice, ANOVA inference is relatively **robust** when the populations are non-Normal, especially when the samples are large and roughly the same size. Before doing the FF test, check the observations in each sample for outliers or strong skewness. Also verify that the largest sample standard deviation is no more than twice as large as the smallest standard deviation.
- When the null hypothesis is true and the conditions for inference are met, the **ANOVA** FF **statistic** for comparing kk means from a total of NTotalN_{Total} observations in all samples combined has the FF **distribution** with k-1$k - 1$ and NTotal-k$N_{Total} - k$ degrees of freedom.
- The *P*-value of the one-way ANOVA is the probability, when H0H_0 is true, of getting a test statistic FF at least as extreme as that obtained, computed as the corresponding right-tail area under the FF distribution with the appropriate degrees of freedom.
- ANOVA calculations are reported in an **ANOVA table** that gives sums of squares, mean squares, and degrees of freedom for variation among groups and for variation within groups. In practice, we use software to do the calculations.

THIS CHAPTER IN CONTEXT

In Chapter 17 we described the one-sample *t* procedures used to perform statistical inference for one population mean μ.μ. This was followed in Chapter 18 by the two-sample *t* procedures used to perform statistical inference for the difference μ1-μ2$\mu_1 - \mu_2$ between the means of two distinct populations, based on two independent random samples. In this chapter, we address how analysis of variance, specifically one-way ANOVA, can be used to compare the means from several populations when we have obtained several independent random samples.

ANOVA is a first step in comparing the means of several populations. To avoid the problems associated with multiple comparisons, we must first ask if there is evidence that these population means are not all equal. This is what the one-way ANOVA tests. When the test is statistically significant, we have evidence that at least one population mean in the set is significantly different from at least one other population mean. ANOVA, however, does not tell us which means differ from which. Companion Chapter 26 describes how we can perform follow-up tests that compare the population means two at a time while compensating for the resulting multiple comparisons.

Analysis of variance is a versatile statistical procedure that can be adapted to more complex settings. Companion Chapter 26 introduces the two-way ANOVA procedure, which allows us to compare population means obtained from more complex designs involving two different factors, or explanatory variables.

The tt procedures and ANOVA are used when the response variable is quantitative. The appropriate procedure to use depends on the study design, something that we discussed in Chapters 6 and 7. Concerns about how the data were collected should lead you to question the extent or even the validity of the conclusions from any statistical inference based on these data. Moreover, the conditions for inference typically require that we examine quantitative data for evidence of skew or outliers, as seen in Chapters 1 and 2.

CHECK YOUR SKILLS

24.12 The purpose of analysis of variance is to compare

a. the variances of several populations.
b. the proportions of successes in several populations.
c. the means of several populations.

24.13 A study of the effects of smoking classifies subjects as nonsmokers, moderate smokers, or heavy smokers. The investigators interview a sample of 200 people in each group. They ask, “How many hours do you sleep on a typical night?” The null hypothesis for the ANOVA FF test is that all three population mean hours of sleep are the same. What is the alternative hypothesis?

a. The mean hours of sleep is smallest in the population of heavy smokers and largest in the population of nonsmokers.
b. The three population mean hours of sleep are all different.
c. The three population mean hours of sleep are not all the same.

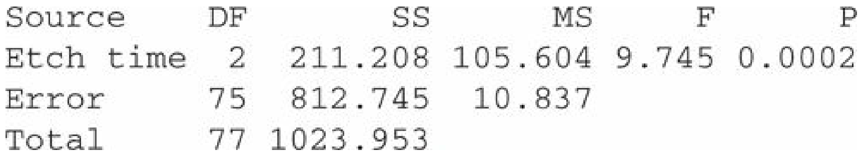
24.14 What are the degrees of freedom for the ANOVA FF statistic comparing mean hours of sleep in Exercise 24.13?

a. 2 and 197
b. 2 and 597
c. 3 and 597

A dental study evaluated the effect of tooth etch time on resin bonding strength. A total of 78 undamaged, recently extracted first molars (baby teeth) were randomly assigned to be etched with phosphoric acid gel for either 15, 30, or 60 seconds. Composite resin cylinders of identical size were then bonded to the tooth enamel. The researchers examined the bond strength after 24 hours by finding the failure load (in megapascals) for each bond. Here are the summary data and ANOVA table for this experiment:[8]

Etch time	**n**n	**x-**$\bar{x}$	**s**s

15 seconds	26	4.49	2.28
30 seconds	26	6.98	3.15
60 seconds	26	8.48	4.17



```
Source      DF        SS       MS      F       P
Etch time    2   211.208 105.604 9.745 0.0002
Error       75   812.745  10.837
Total       77 1023.953
```

Exercises 24.15 to 24.21 are based on this study.

24.15 The most striking conclusion from the numerical summaries for the three etch times is that

a. there appears to be little difference among the etch times.

b. on average, failure load increases with etch time.

c. on average, failure load decreases with etch time.

24.16 If we used a series of two-sample *t* procedures to compare the three conditions, we would have to give three 95% confidence intervals to compare all three pairs of etch times. The weakness of doing this is that

a. we won’t be 95% confident that all 3 intervals cover the true differences in population means.

b. the conclusions from the three intervals might not agree.

c. the conditions for two-sample *t* inference are not met for all 3 pairs of etch times.

24.17 What is the null hypothesis for this ANOVA F*F* test?

a. The population mean load failures for the three etch times are all the same.

b. The population mean load failures for the three etch times are not all the same.

c. The population mean load failure is lowest for the 15-second condition and highest for the 60-second condition.

24.18 What is the alternative hypothesis for this ANOVA F*F* test?

a. The population mean load failures for the three etch times are not all the same.

b. The population mean load failures for the three etch times are all different.

c. The population mean load failure is lowest for the 15-second condition and highest for the 60-second condition.

24.19 The conclusion of the ANOVA test is that

a. there is quite strong evidence (P=0.0002)$(P = 0.0002)$ that the mean failure loads are not the same in all three etch conditions.

b. there is quite strong evidence (P=0.0002)$(P = 0.0002)$ that the mean failure load is much lower for 15 seconds of etch time than for any other two etch times.

c. the data give no evidence (P=0.0002)$(P = 0.0002)$ to suggest that mean failure loads differ among the three etch times.

24.20 After checking the condition that all populations compared have the same standard deviation, you decide that this condition

a. is not met, because the sample standard deviations are too different.

b. is not met, because we don't know the population standard deviations.

c. is met, because the sample standard deviations are similar enough. (The largest one is less than twice the smallest.)

24.21 The researchers stated that the distributions of failure loads were not Normally distributed but rather right-skewed (without extreme outliers). The test *P*-value of 0.0002 is

a. reasonably accurate anyway because the sample sizes are large (26 each).

b. inaccurate because the populations must be Normally distributed to perform ANOVA.

c. unreliable no matter how the data are distributed because these data can't be regarded as random samples or a randomized experiment.

Chapter 24 EXERCISES

24.22 Caffeine and sugar. A double-blind randomized experiment assigned healthy undergraduate students to drink one of four beverages after fasting overnight: water, water with 75 mg of caffeine, water with 75 g of glucose, and water with 75 mg of caffeine and 75 g of glucose. Subjects performed a number of cognitive tasks, including the California Computerized Assessment Package, a computerized reaction time program measuring sustained attention, reaction time, and visual scanning speed. Here are the resulting reaction times (SEM is the standard error of the mean):[9]

Beverage	n	$\bar{x}$	SEM
Water	18	389.35	18.50
Water and caffeine	18	320.16	17.98
Water and glucose	18	318.16	17.04
Water, caffeine, and glucose	18	336.44	14.02

Identify the populations and the response variable. State the null and alternative hypotheses of the corresponding ANOVA F test and specify its degrees of freedom.

24.23 A medical study. The Québec (Canada) Cardiovascular Study recruited men aged 34 to 64 years at random from towns in the Québec City metropolitan area. Of these, 1824 met the criteria (no diabetes, free of heart disease, and so on) for a study of the relationship between being overweight and medical risks. The 719 normal-weight men had a mean triglyceride level of 1.4 millimoles per liter (mmol/l); the 885 overweight men had a mean of 1.7 mmol/l; and the 220 obese men had a mean of 1.9 mmol/l. The sample standard deviations were very similar for the 3 groups.[10] Identify the populations and the response variable. State the null and alternative hypotheses for the corresponding ANOVA F test and specify its degrees of freedom.

24.24 Caffeine and sugar, continued. Go back to Exercise 24.22.

a. Do the conditions for ANOVA appear to be met? (Be sure to consider the relationship between the standard error of the mean and the standard deviation for a sample.)
b. The test statistic and *P*-value for the corresponding ANOVA are F=3.83$F = 3.83$ and P=0.0134.$P = 0.0134$. What can you conclude from this study?
c. Use technology or Table F at the back of the book to obtain the *P*-value from the test statistic provided.

24.25 A medical study, continued. Go back to Exercise 24.23.

a. Do the conditions for ANOVA appear to be met?
b. The ANOVA test statistic is F=12.74,$F = 12.74$, which gives a highly statistically significant *P*-value of 0.000003. Explain why this very small *P*-value does not imply that the mean triglyceride levels in these populations differ a lot.
c. Write a short conclusion in the context.

24.26 Whose little bird is that? The males of many animal species signal their fitness to females by displays of various kinds. For male barn swallows, the signal is the color of their plumage. Does better color really lead to more offspring? A randomized comparative experiment assigned barn swallow pairs that had already produced a clutch of eggs to three groups: 13 males had their color “enhanced” by the investigators; 9 were handled but their color was left alone; and 8 were not touched. DNA testing showed whether the eggs were sired by the male of the pair or by another male. The investigators then removed the eggs. Barn swallows usually produce a second brood if the first one failed. The males with enhanced color fathered more young in the second brood, and males in the other groups fathered fewer. The investigators used ANOVA to compare the 3 groups.[11]

a. What are the degrees of freedom for the FF statistic that compares the number of young sired by each male among the eggs in the first brood? The statistic was F=0.07.$F = 0.07$. What can you say about the *P*-value? What do you conclude?

b. Of the 30 original pairs, 27 produced a second brood. What are the degrees of freedom for FF to compare differences in the number of young sired by each male between the first and second broods produced by these 27 pairs? The statistic was F=5.45.$F = 5.45.$ What can you say about the *P*-value? What do you conclude?

24.27 Plants defend themselves. When some plants are attacked by leaf-eating insects, they release chemical compounds that both repel the leaf-eaters and attract other insects that prey on the leaf-eaters. An experiment examined the release of these protective compounds by a plant that grows in the Utah desert.[12] The investigators compared plants attacked by one of three types of leaf-eaters and other plants that were undamaged—32 plants of the same species in all, 8 per group. They then measured the emission rate of a repelling compound during seven hours. Here are the data (mean ±$\pm$ standard error of the mean), measured in nanograms per hour (ng/hr).

Group	**Emission rate (ng/hr)**
Control	9.22±5.93$9.22 \pm 5.93$
Hornworm	31.03±8.75$31.03 \pm 8.75$
Leaf bug	18.97±6.64$18.97 \pm 6.64$
Flea beetle	27.12±8.62$27.12 \pm 8.62$

a. Create a graph that compares the mean emission rates for the four groups. Does it appear that emissions increase when the plant is attacked?

b. What hypotheses does ANOVA test in this setting?

c. We do not have the full data set. What would you look for in deciding whether you can safely use ANOVA?

d. What is the relationship between the standard error of the mean and the standard deviation for a sample? Do the standard deviations satisfy our rule of thumb for safe use of ANOVA?

e. The test *P*-value is 0.207. What do you conclude?

24.28 Logging in the rainforest: species counts. Exercise 24.5 described a study of rainforest recovery after logging. Use software to analyze the effect of logging on the number of species per plot, provided in Table 24.2.

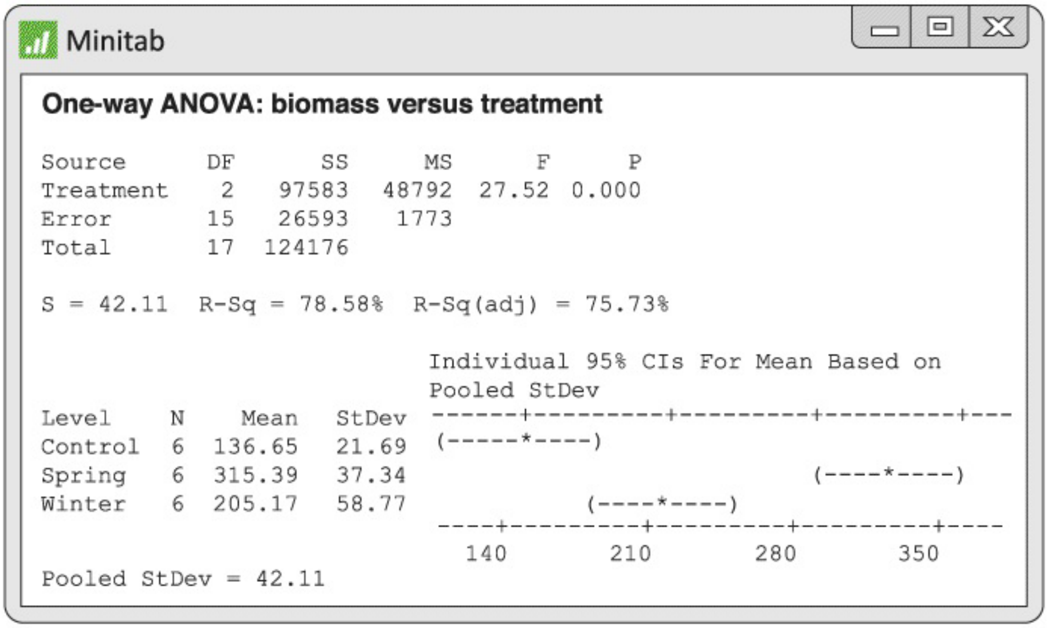
a. Plot the data and report the group means and standard deviations. What do the means suggest about the effect of logging on the number of tree species? Are the conditions for inference with ANOVA met?
b. Carry out the ANOVA. Report the FF statistic and its *P*-value and conclude in context.

24.29 More rain for California? The changing climate will probably bring more rain to California, but we don't know whether the additional rain will come during the winter wet season or extend into the long dry season in spring and summer. Kenwyn Suttle of the University of California at Berkeley and his coworkers randomly assigned plots of open grassland to three treatments: added water equal to 20% of annual rainfall either during January to March (winter) or during April to June (spring), and no added water (control). Here are some of the data, for plant biomass (in grams per square meter) produced by each plot in a single year.[13]

Winter	**Spring**	**Control**
264.1514	318.4182	129.0538
187.7312	281.6830	144.6578
291.1431	288.8433	172.7772
176.2879	382.6673	113.2813
141.7525	326.8877	142.1562
169.9737	293.8502	117.9808

Figure 24.7 shows Minitab ANOVA output for these data.

a. Plot the data and obtain the sample means and standard deviations. What do the data appear to show about the effect of extra water in winter and in spring on biomass? Do these data satisfy the conditions for ANOVA?
b. State H0H_0 and HaH_a for the ANOVA FF test, and explain in words what ANOVA tests in this setting.
c. Report your overall conclusions about the effect of added water on plant growth in California. Be sure to report the test statistic and *P*-value.



One-way ANOVA: biomass versus treatment

```
Source      DF      SS     MS      F      P
Treatment    2   97583  48792  27.52  0.000
Error       15   26593   1773
Total       17  124176

S = 42.11  R-Sq = 78.58%  R-Sq(adj) = 75.73%

                            Individual 95% CIs For Mean Based on
                            Pooled StDev
Level    N    Mean   StDev  ------+---------+---------+---------+---
Control  6  136.65   21.69  (-----*----)
Spring   6  315.39   37.34                             (----*----)
Winter   6  205.17   58.77             (----*----)
                             ----+---------+---------+---------+----
                               140       210       280       350

Pooled StDev = 42.11
```

Figure 24.7

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 24.7 Minitab ANOVA output for comparing the total plant biomass of grassland plots under different water conditions.

24.30 Naloxone. Nalaxone is an opioid blocker critical to the treatment of opiod overdose, but its mechanism of action is not yet fully understood. Arachidonyl-29-chloroethylamide (ACEA) is a molecule designed to activate a specific cannabinoid receptor in the brain so that the receptor's function can be studied. Like cannabis, ACEA induces a hyperfeeding response in mice. Researchers randomly assigned 22 mice that had just fed to 4 treatments consisting of an injection: ACEA alone, nalaxone alone, ACEA plus nalaxone, saline only (control). They then measured the amount of food the mice consumed within 1 hour of treatment. Here are the findings (in milligrams, mg):[14]

ACEA	Naloxone	ACEA + naloxone	Control
130	10	10	20
150	40	50	30
150	40	60	40
190	120	80	80
250	120	120	170
	140	160	

What does your analysis suggest about the involvement of naloxone in cannabinoid-induced hyperfeeding in mice? Use the four-step process to guide your discussion.

24.31 Evolution in bacteria. Can bacteria evolve a preference for the pH of their environment? An evolutionary biologist took lines of *Escherichia coli* bacteria grown and kept at a neutral pH of 7.2 and grew them for 2000 generations (about 300 days) at a stressful acidic pH of 5.5. The original bacteria (kept frozen all that time) and the "acid-evolved" bacteria were then grown together in various test environments, and a relative fitness score was computed. A score of 1 would indicate that both bacteria types were equally fit. A score higher than 1 would indicate that the acid-evolved line was more fit than the ancestral line in that environment (that is, that the acid-evolved bacteria grew the most). Here are the relative fitness scores obtained for test environments of pH 5.5 (acid), 7.2 (neutral), or 8.0 (basic). There are six replicates using different ancestor bacterial lines for each test environment:[15]

Test in acid pH	1.24	1.22	1.23	1.24	1.18	1.09
Test in neutral pH	0.99	0.99	0.98	0.94	0.95	0.95
Test in basic pH	0.56	0.83	0.82	0.72	0.86	0.84

Do the data provide evidence that *E. coli* can evolve adaptations to the pH of its environment over the course of 2000 generations? Follow the four-step process in answering this question. (Although the standard deviations do not quite satisfy our rule of thumb, that rule is conservative and small sample sizes can more easily lead to widely different variances. The analysis might be performed again without the low outlier to confirm the initial interpretation.)

24.32 Neural basis of 3D vision. One way the brain assesses depth in a visual scene is by using the disparities in the information received by both retinas. Stereograms, like 3D movies, contain a visual pattern that is duplicated and shifted a bit horizontally. Each pattern is seen by just one eye through the use of special 3D glasses, introducing retinal disparity (measured in degrees). The result is an image that appears to float either in front of or behind the screen, depending on the horizontal disparity introduced.

Some neurons in the primary visual cortex specialize in the perception of depth from binocular disparity. One such neuron was recorded in a monkey shown stereograms of various disparities in random order. Figure 24.8(a) shows the action potentials (“spikes”) generated over time (horizontal axis) by that neuron for different disparities. Figure 24.8(b) shows the “tuning curve” representing the mean neural response (±SEM) (±SEM) for each disparity. Here are the raw data from 14 replications:[16]

Disparity	Neural response to stereogram (spikes/s)													
− 0.6	10	46.7	83.3	80	40	36.7	46.7	46.7	70	76.7	20	53.3	66.7	83.3
− 0.3	30	110	100	90	70	56.7	90	90	130	103.3	60	86.7	126.7	123.3
− 0.15	70	93.3	86.7	76.7	103.3	53.3	106.7	120	100	126.7	30	100	113.3	96.7
0	10	13.3	36.7	16.7	36.7	16.7	20	23.3	23.3	16.7	26.7	23.3	13.3	36.7
0.15	13.3	63.3	10	63.3	93.3	16.7	10	26.7	23.3	70	3.3	23.3	20	40
0.3	6.7	23.3	50	43.3	76.7	10	46.7	83.3	60	26.7	16.7	30	46.7	63.3

A negative disparity corresponds to the impression of a near object. This neuron clearly responds more strongly to objects that appear in the near space. Does your analysis support this interpretation? Use the four-step process to guide your discussion.

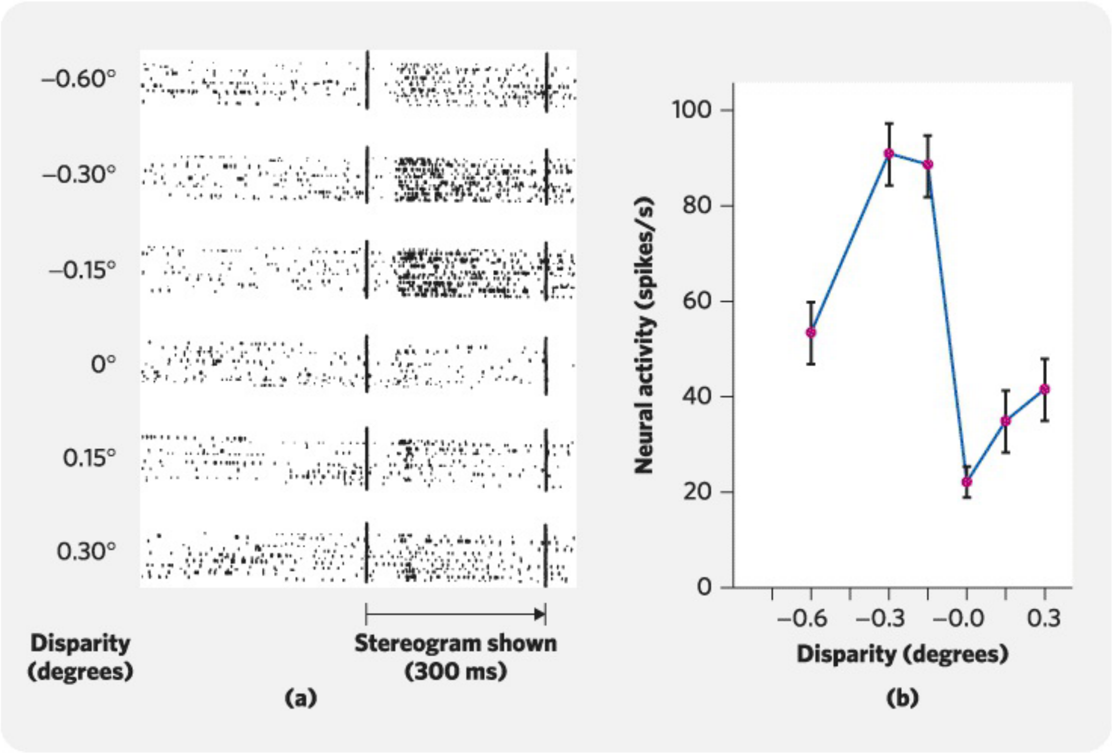


Figure 24.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 24.8 (a) Action potentials (dots) over time in response to 3D visual stimuli of different disparities. (b) Graph of the mean response ± standard error of the mean for each disparity value.

24.33 Fertilizing bromeliads. Bromeliads are tropical flowering plants. Many are epiphytes that attach to trees and obtain moisture and nutrients from air and rain. Their leaf bases form cups that collect water and are home to the larvae of many insects. As a preliminary to a study of changes in the nutrient cycle, Jacqueline Ngai and Diane Srivastava examined the effects of adding nitrogen, phosphorus, or both to the cups. They randomly assigned 8 bromeliads growing in Costa Rica to each of 4 treatment groups, including an unfertilized control group. A monkey destroyed one of the plants in the control group, leaving 7 bromeliads in that group. Here are the numbers of new leaves on each plant over the seven months following fertilization:[17]

Nitrogen	Phosphorus	Both	Neither
15	14	14	11
14	14	16	13

15	14	15	16
16	11	14	15
17	13	14	15
18	12	13	11
17	15	17	12
13	15	14	

Does nitrogen or phosphorus have a greater effect on the growth of bromeliads? Analyze these data and discuss the results.

24.34 Compressing soil. Farmers know that driving heavy equipment on wet soil compresses the soil and injures future crops. Table 24.3 gives data on the "penetrability" of the same soil at three levels of compression.[18] Penetrability is a measure of how much resistance plant roots will meet when they try to grow through the soil. Do the data suggest that penetrability decreases as soil is more compressed? Follow the four-step process in data analysis and ANOVA. Be sure to check the conditions for ANOVA, paying special attention to outliers.

TABLE 24.3 Penetrability of soil at three compression levels

Compressed	2.86	2.68	2.92	2.82	2.76	2.81	2.78	3.08	2.94	2.86
	3.08	2.82	2.78	2.98	3.00	2.78	2.96	2.90	3.18	3.16
Intermediate	3.13	3.38	3.10	3.40	3.38	3.14	3.18	3.26	2.96	3.02
	3.54	3.36	3.18	3.12	3.86	2.92	3.46	3.44	3.62	4.26
Loose	3.99	4.20	3.94	4.16	4.29	4.19	4.13	4.41	3.98	4.41
	4.11	4.30	3.96	4.03	4.89	4.12	4.00	4.34	4.27	4.91

24.35 Homeopathic treatment of wounds. Homeopathy seeks to stimulate the body's ability to heal itself by giving very small doses of highly diluted and generally harmful substances. The dilution can be so high that no molecules of the healing substance remain in any traceable amount. (Homeopathy proposes that the substance leaves an imprint called the "memory of water.") The National Center for Complementary and Alternative Medicine (NCCAM) warns that, except for a few studies, there is little evidence to support homeopathy as an effective treatment for any specific condition.

In one study, researchers sutured a deep muscular incision that they had made in anesthetized rats and then randomly assigned the rats to 5 treatments: a placebo (sugar water) and two homeopathic remedies (*Arnica montana* and *Staphisagria*) administered at two levels of dilution. The treatments were injected daily into the rats' stomachs via a gastric tube, and the wounds were examined daily to determine the time (in days) until complete healing of the wound. Table 24.4 gives the raw data.[19] Is there significant evidence that healing time depends on the treatment received? Use the four-step process to guide your discussion.

TABLE 24.4 Healing times (in days) for a deep surgical wound

Placebo	Arnica low dose	Arnica higher dose	Staphisagria low dose	Staphisagria higher dose
21	16	16	15	16
22	15	16	15	16
22	16	16	16	16
21	15	16	16	15
22	15	15	15	15
21	16	15	16	16
20	16	16	15	16
22	15	16	15	16
21	15	16	16	15
21	16	15	16	16
22	15	16	16	15
23	16	16	15	16
21	16	15	15	15
22	15	16	16	15
21	15	15	15	16

24.36 Did the randomization work? When assigning patients with existing symptoms to various treatments in a clinical trial, it is important to verify that randomization did not, by chance, create uneven groups, with different overall levels of symptom severity. A study of low bone mineral density (BMD) among postmenopausal women randomly assigned subjects to 5 treatment groups. BMD at the lumbar spine was measured at the beginning of the study to assess symptom severity. BMD is expressed as a unitless, standardized value; a negative score indicates a BMD less than the normal healthy average for a population of young women. Here are the summarized BMD scores for the 5 groups at the beginning of the study:[20]

Treatment group	**n**	**$\bar{x}$**	**s**
Placebo	46	−2.2	0.7
Denosumab, 6 mg	44	−2.0	0.9
Denosumab, 14 mg	44	−2.0	0.8
Denosumab, 30 mg	41	−2.2	0.7
Alendronate	47	−2.0	0.9

Calculate the ANOVA table and the F statistic. Are the differences in mean BMD scores among the 5 groups significant? What does that say about the randomization process for this study?

24.37 Plant defenses (optional). ANOVA calculations use only the sample sizes n_i, the sample means $\bar{x}_i$, and the sample standard deviations s_i. You can therefore recreate the ANOVA calculations when a report gives these summaries but does not give the actual data. Use the information in Exercise 24.27 to calculate the ANOVA table (sums of squares, degrees of freedom, mean squares, and the F statistic). Note that the report gives the standard error of the mean rather than the standard deviation. Are there significant differences among the mean emission rates for the four populations of plants?

24.38 F versus t (optional). Acclimation is a process of adaptation to a persistent change in environment taking place typically over the course of weeks. When goldfish are relocated to a colder aquarium, for instance, their ventilation rate immediately slows down but, after several weeks, moves back toward its previous value. This acclimation process can be complete (the rates before and after acclimation are exactly the same) or only partial. Here are ventilation rates (in beats per minute) for goldfish that have been acclimated to either a cold environment (12 degrees Celsius) or a warm environment (22 degrees Celsius):[21]

Group	nn	x̄$\bar{x}$	ss
Cold-acclimated	18	55.8	18.4
Warm-acclimated	18	78.3	25.8

A significant difference in ventilation rate between warm- and cold-acclimated goldfish would indicate that acclimation was not complete.

a. Calculate the two-sample *t* statistic for testing H0: μ1=μ2$H_0 : \mu_1 = \mu_2$ against the two-sided alternative. Use the conservative method to find the *P*-value.
b. Calculate MSG, MSE, and the ANOVA FF statistic for the same hypotheses. What is the *P*-value of F?$F?$
c. How close are these two *P*-values?

24.39 Elderly health. A study examined the medical records of elderly patients for calcium or inorganic phosphorus blood levels (both in mmol/l). The data file *Large.Calcium*, which you examined in Exercise 18.43, contains the data. One purpose of the study was to determine if different laboratories yield substantially different results. Do the data support this hypothesis? Verify that the conditions for ANOVA are met for the calcium variable, run the test, and conclude. Then do the same with the inorganic phosphorus data. Write a short paragraph describing your findings.

24.40 The cost of sex. Evolution theory predicts a physiological cost of increased reproduction that translates into a shorter life span. The data file *Large.Fruitfly* describes a study of the cost of sexual activity in male *Drosophila*. The authors randomly assigned male fruit flies to 5 sex conditions. In the first condition, the males were kept alone. In the second condition, each male was placed in the company of one virgin female (replaced every other day with a new virgin female to provide reproductive opportunities). The third condition had each male placed with eight virgin females. In the fourth condition, each male was placed with a newly inseminated female that was replaced every other day; these females do not usually mate for at least two days. The fifth condition had each male placed in the company of eight newly inseminated females. All other living conditions were the same for all five groups. The researchers recorded the longevity (in days) of each male fruit fly and its thorax length.[22]

a. The researchers confirmed by examining the females that the males kept with virgin females were indeed reproducing and that those kept with newly inseminated females were not. Is there evidence that the different sex conditions result in significantly different longevities in laboratory male fruit flies? Check the conditions for ANOVA, run the test, and conclude.
b. Is there evidence that the different sex conditions result in significantly different thorax lengths in laboratory male fruit flies? Check the conditions for ANOVA, run the test, and conclude.
c. Write a short paragraph describing your findings.

CHAPTER 25 Statistical Inference: Part III Review

Linnette Engler/Shutterstock.com

IN THIS CHAPTER WE COVER...

- Essential Concepts from Parts I and II
- Part III Summary
- Comprehensive Review Exercises: Inference Selection
- Comprehensive Review Exercises: Analysis and Conclusion
- Large Data Set Exercises
- Online Data Sources
- EESEE Case Studies

The procedures covered in Chapters 17 through 24 are among the most commonly used of all statistical inference methods. Now that you have mastered important ideas and practical methods for inference, it’s time to review the big ideas of statistics in outline form. The following selective summary of Parts I and II of this book outlines the concepts that are most directly relevant to statistical inference. The outline contains important warnings: Look for the Caution icon.

ESSENTIAL CONCEPTS FROM PARTS I AND II

A. Data Production

- Data basics:
  - Individuals (units, subjects).
  - Variables: categorical versus quantitative; explanatory versus response.
  - Purpose of study.
- Data production basics:
  - Observation versus experiment.
  - Simple random samples.
  - Comparative randomized experiments.
  - Lurking variables.
- Beware: *Really bad data production (voluntary response or confounding) can make interpretation impossible.*
- Beware: *Weaknesses in data production (for example, sampling students at only one campus) can make generalizing conclusions difficult.*

B. Data Analysis

- Always plot your data. Look for an overall pattern and striking deviations.
- Add numerical descriptions based on what you see.
- *Beware the effects of outliers and influential observations.*
- Beware: *Averages and other simple descriptions can miss the real story.*
- One quantitative variable:
  - Graphs: dotplot, histogram, time plot, boxplot.
  - Pattern: distribution shape, center, spread. Outliers?
  - Numerical descriptions: five-number summary or x-$\bar{x}$ and s.s.
  - Compare the distribution of a quantitative variable among different groups to describe the relationship between one quantitative variable and one categorical variable (groups).

- Relationships between two quantitative variables:
  - Graph: scatterplot.
  - Pattern: relationship form, direction, strength. Outliers? Influential observations?
  - Numerical descriptions for linear relationships: correlation, regression line.
  - *Beware the lurking variable: Correlation does not imply causation.*
- Relationships between two categorical variables:
  - Two-way table of counts.
  - Numerical descriptions for each variable: marginal distributions (row percents, column percents).
  - Numerical descriptions for the relationship between the two factors: conditional distributions.
  - Beware: *A lurking variable could create Simpson's paradox.*

C. The Reasoning of Statistical Inference

- Statistical inference uses sample data to infer conclusions about a wider population.
- When we do inference, we are acting as if the data come from random samples or randomized comparative experiments. Beware: *If they don't, we may have "garbage in, garbage out."*
- Always examine your data before doing inference. Inference for quantitative data often requires a regular pattern, such as roughly Normal with no strong outliers.
- Confidence intervals: Estimate a population parameter.
  - 95% confidence: We used a method that captures the true parameter 95% of the time in repeated use.
  - Beware: *The margin of error of a confidence interval does not include the effects of practical errors such as undercoverage and nonresponse.*
- Hypothesis tests: Assess the evidence against H0H_0 in favor of Ha. H_a.
  - *P*-value: If the test assumptions are met and H0H_0 was true, how

often would we get an outcome favoring the alternative this strongly? Smaller *P* is stronger evidence against H0.H_0.

– Statistical significance at the 5% level, P<0.05,$P < 0.05$, means that an outcome this extreme would occur by chance less than 5% of the time if H0H_0 was true.

– Beware: *The value 0.05 is not extraordinary, so use* P<0.05$P < 0.05$ *more as a safeguard against unfair conclusions than as an arbitrary threshold to reject or fail to reject* H0.H_0.

– Sample size affects *P*-value computations. Large samples can make even small effects statistically significant. Small samples can fail to identify even large effects as statistically significant.

– Beware: *Statistical significance is not the same as practical importance.* Assessing practical importance requires expert knowledge of the context.

– Always try to estimate the size of an effect (for example, with a confidence interval), not just its statistical significance.

PART III SUMMARY

Chapters 17 through 24 introduced a number of different inference procedures. (Many more exist, some of which are covered in optional Chapters 26 through 28 on the companion website.) Because the computations required for inference are now typically automated with technology, two aspects of statistical inference still require critical thinking: Selecting an appropriate inference procedure and checking the conditions for inference. Students often find selecting an inference procedure to be the most challenging. Exercises 25.1 through 25.9 help you practice this skill. Choose inference procedures by asking:

- “What parameter?” and “What study design?” Matched pairs (repeated measures) and independent samples require different computational approaches.
- “How many variables?” and “How many groups compared?”
- “What kind of variables?” Categorical and quantitative variables need very different inference procedures.

The flowcharts in Figures 25.1 and 25.2 help you decide. The flowchart for means and proportions (Figure 25.1) is based on the first two questions and the flowchart for relationships (Figure 25.2) is based on the third:

- **What population parameter** does your problem concern? You should be familiar in detail with inference for *means* (Chapters 17, 18, and 24) and for *proportions* (Chapters 19, 20, 21, and 22). This is the first branch point in the flowchart of Figure 25.1.
- **What type of design** produced the data? You should be familiar with data from a *single sample*, from *matched pairs*, and from *two or more independent samples.* This is the second branch point in Figure 25.1. Your choice here is the same for experiments and for observational studies—inference methods are the same for both types of study. But don’t forget that experiments give much better evidence that an effect uncovered by inference reflects direct

causation.

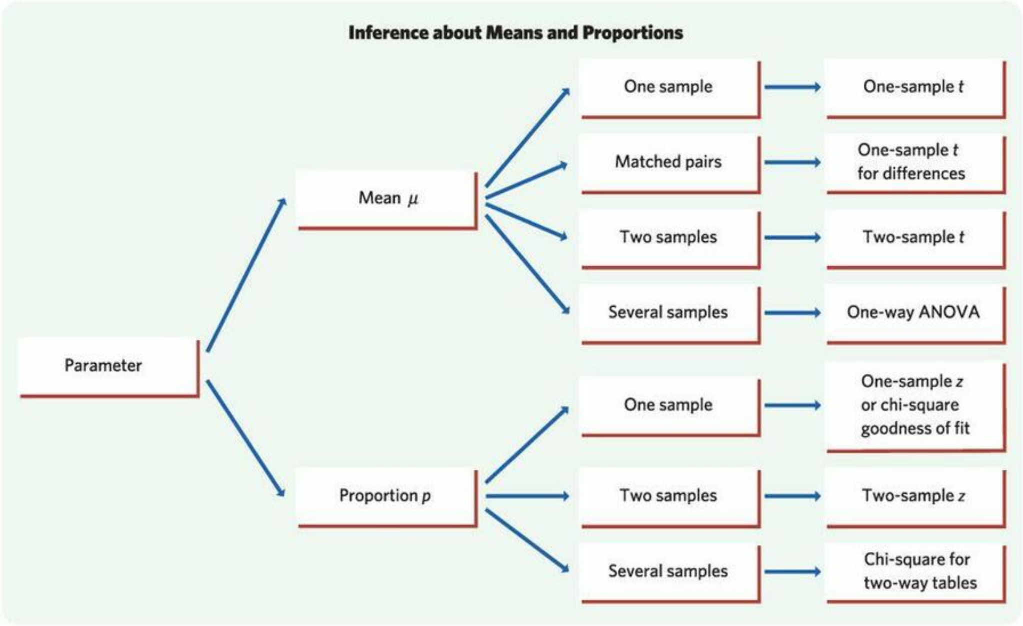

Inference about Means and Proportions
Parameter
Mean μ
Proportion p
One sample
Matched pairs
Two samples
Several samples
One sample
Two samples
Several samples
One-sample t
One-sample t for differences
Two-sample t
One-way ANOVA
One-sample z or chi-square goodness of fit
Two-sample z
Chi-square for two-way tables

Figure 25.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 25.1

- **What type of variables** make up the relationship when studying patterns of association? The choices are illustrated in Figure 25.2. You should be familiar in detail with inference for relationships between *two categorical variables* (Chapter 22), between *two quantitative variables* (Chapter 23), and between *a quantitative variable and a categorical variable* (Chapter 24).

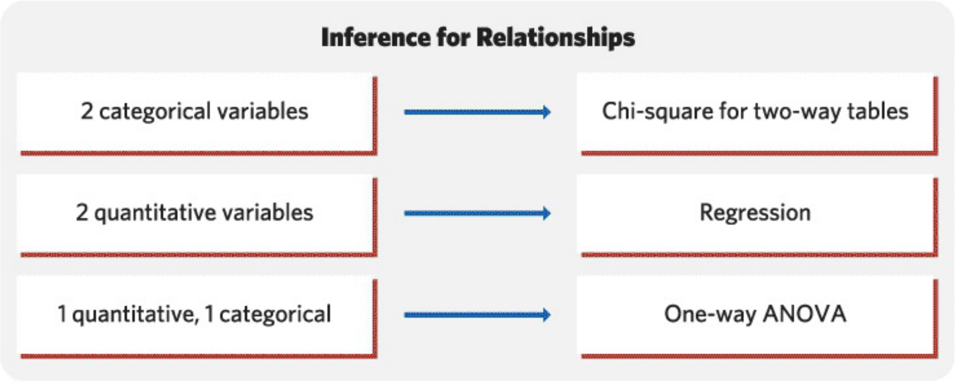

Inference for Relationships
2 categorical variables
Chi-square for two-way tables
2 quantitative variables
Regression
1 quantitative, 1 categorical
One-way ANOVA

Figure 25.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 25.2

Answering these questions is part of the Plan step in the four-step process. To begin the Solve step, follow the flowcharts to the correct procedure. Notice that some procedures (ANOVA, chi-square for two-way tables) appear in both flowcharts. This is mostly a matter of perspective. When comparing the distribution of one variable among several groups, the groups themselves form a categorical explanatory variable, which means that the approach could also be considered as inference for a relationship between an explanatory and a response variable.

The other critical part of statistical inference that cannot be automated with technology is checking the conditions for inference. This, too, is part of the Solve step in the four-step process. Ask:

- **Are the conditions for this procedure met?** Can you act as if the data come from a random sample or a randomized comparative experiment? With quantitative variables, for instance, do the data show extreme outliers or strong skewness that rules out use of inference based on Normality? Do you have enough observations to rely on the robustness of your intended procedure? And so on.

Mastery of statistics comes with practicing. The review exercises in this chapter allow you to decide for yourself which of several inference procedures to use. Learning to recognize problem settings so as to choose the right type of inference is a key step in advancing your mastery of statistics.

Here are the most important skills you should have acquired from reading Chapters 17 through 24.

A. Recognition

1. Recognize when a problem requires inference about population means (quantitative response variable), population proportions (categorical response variable), or relationships.
2. Recognize from the design of a study whether one-sample, matched pairs, or two-sample procedures for one variable are needed, or whether the design assesses the relationship between two variables.
3. For one-variable problems, choose among the one- and two-sample t procedures for means, the one- and two-sample z procedures for

proportions, and the chi-square goodness-of-fit test for proportions. For relationships, choose among the chi-square procedures for two categorical variables, regression procedures for two quantitative variables, and the ANOVA FF procedures for a quantitative and a categorical variable.

B. Inference About One Mean

1. Verify that the *t* procedures are appropriate in a particular setting. Check the study design and the distribution of the data, and take advantage of robustness against lack of Normality for large enough sample sizes.
2. Recognize when poor study design, outliers, or a small sample from a skewed distribution make the *t* procedures risky.
3. Use the one-sample *t* procedure to obtain a confidence interval at a stated level of confidence for the mean μμ of a population.
4. Carry out a one-sample *t* test for the hypothesis that a population mean μμ has a specified value against either a one-sided or a two-sided alternative. Obtain the *P*-value and conclude in context.
5. Recognize matched pairs data and use the *t* procedures to obtain a confidence interval and to perform a hypothesis test for such data.

C. Comparing Two Means

1. Verify that the two-sample *t* procedures are appropriate in a particular setting. Check the study design and the distribution of the data, and take advantage of robustness against lack of Normality for large enough sample sizes.
2. Use the two-sample *t* procedure to obtain a confidence interval at a stated level of confidence for the difference between two population means.
3. Carry out a two-sample *t* test for the hypothesis that two populations have equal means against either a one-sided or a two-sided alternative. Obtain the *P*-value and conclude in context.
4. Know that procedures for comparing the standard deviations of two Normal populations are available, but that these procedures are risky because they are not at all robust against non-Normal distributions.

D. Inference About One Proportion

1. Verify that you can safely use either the large-sample or the plus four zz procedures in a particular setting. Check the study design and the guidelines for sample size and counts of successes and failures.

2. Use the large-sample zz procedure to obtain a confidence interval for a population proportion p.p. Understand that the true confidence level may be substantially less than you ask for unless the sample is very large and the true value of pp is not close to 0 or 1.

3. Use the plus four modification of the zz procedure to obtain a confidence interval for pp that is accurate even for small samples and for any value of p.p.

4. Carry out a one-sample zz test for the hypothesis that a population proportion pp has a specified value against either a one-sided or a two-sided alternative. Obtain the *P*-value and conclude in context.

E. Comparing Two Proportions

1. Verify that you can safely use either the large-sample or the plus four zz procedures in a particular setting. Check the study design and the guidelines for sample sizes and counts of successes and failures.

2. Use the large-sample zz procedure to obtain a confidence interval for the difference p1-p2$p_1 - p_2$ between proportions in two populations based on independent samples from the populations. Understand that the true confidence level may be less than you ask for unless the samples are quite large.

3. Use the plus four modification of the zz procedure to obtain a confidence interval for p1-p2$p_1 - p_2$ that is accurate even for very small samples and for any values of p1p_1 and p2.p_2.

4. Carry out a two-sample zz test for the hypothesis that the proportions p1p_1 and p2p_2 in two distinct populations are equal against either a one-sided or a two-sided alternative. Obtain the *P*-value and conclude in context.

F. Chi-Square Test for Goodness of Fit

1. State the null hypothesis of your specific chi-square for goodness of fit test, as defined by your study’s objective.

2. Carry out a chi-square test for goodness of fit to test whether a categorical variable has the distribution specified in the null hypothesis against a nonspecific alternative. Obtain the *P*-value and conclude in context.
3. Know that expected counts for a goodness-of-fit test are based on the null hypothesis and sample size. Use the expected counts to check whether you can safely use the chi-square test.
4. If the test is statistically significant, compare percents, compare observed with expected counts, or look for the largest components of the chi-square statistic to see which deviations from the null hypothesis are most important.

G. Chi-Square Test for Two-Way Tables

1. Know how to organize data representing two categorical variables as a two-way table of counts of outcomes.
2. Use percents to describe the relationship between any two categorical variables, starting from the counts in a two-way table.
3. Carry out a chi-square test for two-way tables to test the null hypothesis of no relationship between two categorical variables against a nonspecific alternative of some relationship. Obtain the *P*-value and conclude in context.
4. Know that expected counts for a two-way test are based on the observed counts displayed in the margin of the two-way table. Use the expected counts to check whether you can safely use the chi-square test.
5. If the test is statistically significant, compare percents, compare observed with expected cell counts, or look for the largest components of the chi-square statistic to see which deviations from the null hypothesis are most important.

H. Inference for Linear Regression

1. Create a scatterplot to display the relationship between a quantitative explanatory variable xx and a quantitative response variable y.y. Recognize the regression setting: a straight-line relationship between xx and y.y.

2. Inspect the data to recognize situations in which inference for linear regression isn’t safe: a nonlinear relationship, influential observations, strongly skewed residuals in a small sample, or nonconstant variation of the data points about the regression line.

3. Explain in context the meaning of the slope β of the population regression line.

4. Use a *t* procedure to obtain a confidence interval for β or to test the null hypothesis that $\beta = 0$ against either a one-sided or a two-sided alternative.

5. Explain the distinction between a confidence interval for the mean response and a prediction interval for an individual response.

6. Use a *t* procedure to obtain a confidence interval for the mean response μ_y or a prediction interval for an individual response y when the response variable is equal to a specific value within the range used to obtain the regression model.

I. Comparing Several Means with ANOVA

1. Recognize when you can safely use ANOVA to compare the means of several populations. Check the data production, the presence of outliers, and the sample standard deviations for the groups you want to compare. Take advantage of robustness against lack of Normality for large enough sample sizes.

2. Carry out an ANOVA F test of the null hypothesis that the means of several populations are all equal against a nonspecific alternative that not all the population means are equal. Obtain the *P*-value and conclude in context.

3. If the test is statistically significant, use graphs and descriptive statistics to see which differences among the means are most important (optional Chapter 26 on the companion website covers the details of some common follow-up tests used when the F test is significant).

COMPREHENSIVE REVIEW EXERCISES: INFERENCE SELECTION

These exercises are intended to help you consider which of the statistical inference procedures presented in Chapters 17 through 24 is suited to a particular type of data, study design, and stated objective.

25.1 Dark chocolate treatment. Dark chocolate can dilate arteries by down-regulating oxidative stress in the body. A study investigated whether dark chocolate improves walking autonomy in elderly patients with peripheral artery disease (PAD). The study compared the maximum walking distance (in meters) for 20 such patients once after they ate 40 g of dark chocolate and once after they ate 40 g of milk chocolate. The milk and dark chocolate conditions were administered on two different weeks, chosen at random for each patient.[1] Identify the individuals and variable(s) studied. State the null and alternative hypotheses corresponding to the study's objective and indicate which inference procedure from Chapters 17 through 24 you would consider to test them.

25.2 Misdiagnosed ADHD? Attention-deficit/hyper-activity disorder (ADHD) is a disorder that makes it difficult for a person to pay attention and control impulsive behaviors. Symptoms begin in childhood and the disorder is typically diagnosed at a young age. Increasing rates of ADHD diagnosis worldwide have raised concerns of wrong diagnosis. For instance, the youngest children in the classroom could appear more unruly than their older peers and be misdiagnosed with ADHD. In Taiwan, the cut-off for admission to a given grade level is a birthday later than August 31. Researchers selected a very large random sample of Taiwanese school-age children and compared the proportion of children receiving ADHD medication among those born in August and among those born in September.[2] Identify the individuals and variable(s) studied. State the null and alternative hypotheses corresponding to the study's objective and indicate which inference procedure from Chapters 17 through 24 you would consider to test them.

25.3 Effectiveness of metformin. A study examined the effect of the drug metformin when used in addition to insulin in the treatment of overweight adolescents with type 1 diabetes. The study recruited 140 overweight adolescents with type 1 diabetes treated with insulin. At random, 71 were given metformin and 69 were given a placebo to take in addition to their standard treatment. Six months later, the researchers compared the mean body mass index (in kg/m^2) in the two groups.[3] Identify the individuals and variable(s) studied. State the null and alternative hypotheses corresponding to the study's objective and indicate which inference procedure from Chapters 17 through 24 you would consider to test them.

25.4 Ant choices. Most people favor their right hand for complex tasks, but lateralization is not specific to humans. Researchers studied whether ants show a directional preference when entering a maze. Of 66 ants that showed a consistent preference in repeat maze entries, 29 preferred going left and 37 preferred going right.[4] Identify the individuals and variable(s) studied. State the null and alternative hypotheses corresponding to the study's objective and indicate which inference procedure from Chapters 17 through 24 you would consider to test them.

25.5 BPA in stored water bottles. Bisphenol A (BPA) is an industrial chemical used in the manufacture of some types of plastics, such as disposable water bottles. BPA can mimic the action of estrogen (a human sex hormone), so it is important to know how much may leak out of the plastic and enter into food and drinks. Researchers measured the BPA concentration in bottled water stored for varying amounts of time to assess this relationship.[5] Identify the individuals and variable(s) studied. State the null and alternative hypotheses corresponding to the study's objective and indicate which inference procedure from Chapters 17 through 24 you would consider to test them.

25.6 Concussion and sleep. A 2016 National Public Radio (NPR) report described a study comparing patients who had their first ever traumatic brain injury 18 months earlier with a control group of unrelated healthy individuals similar in age but with no prior brain trauma.[6]

a. The news report stated that,18 months after their traumatic brain injury, the patients "were still getting, on average, an hour more sleep each night than similar healthy people were getting." Which procedure may be used to investigate the statistical significance of this statement?

b. The report further stated that, among those who had experienced brain trauma, "67 percent showed signs of excessive daytime sleepiness. Only 19 percent of healthy people had that problem." Which procedure may be used to investigate the statistical significance of this statement?

c. Can we conclude from this study that trauma to the brain causes sleep disturbances 18 months later? Explain your reasoning.

25.7 A treatment for Zika. Some individuals infected with the Zika virus (ZIKV) develop a neurological disorder called Guillain-Barré syndrome. In addition, a substantial proportion of infants born to mothers who were infected with ZIKV during pregnancy present with severe birth defects, including microcephaly (incomplete neurological development that manifests in an abnormally small brain). A research team investigated the effectiveness of the antiviral drug 25-hydroxycholesterol (25HC) against ZIKV. Some of the findings are reported here.[7] For each part of the study, determine which inference procedure from Chapters 17 through 24 you would consider and state the corresponding null and alternative hypotheses.

a. To investigate the role of 25HC in anti-ZIKV immunity, cell cultures were injected with one of four solutions (low, medium, or high concentration of 25HC, or a placebo) before exposing the cell cultures to ZIKV. The study found that the quantity of ZIKV produced by the cell cultures differed significantly, with a much higher quantity of ZIKV produced in the placebo condition.

b. To assess the protective effect of 25HC on the embryo, the researchers gave pregnant mice, at random, either 25HC or a placebo before infecting them with ZIKV. The study found that the number of ZIKV-infected cells was significantly lower in the cerebral cortex of embryos whose mothers had received 25HC than in the placebo group.

c. To determine whether 25HC could protect against Zika-induced microcephaly in newborns, the researchers gave pregnant mice, at random, either 25HC or a placebo before infecting them with ZIKV.

Treatment with 25HC was found to significantly reduce the percent of newborn mice with microcephaly.

25.8 Tattoos. Permanent tattoos are fairly common but adverse reactions to them are not well documented. Researchers administered a survey about adverse tattoo reactions to a random sample of 300 American adults who had at least one tattoo. Some of the findings are reported here.[8] For each finding, determine which inference procedure from Chapters 17 through 24 you would consider. Also state the corresponding null and alternative hypotheses or specify the parameter estimated by the confidence interval.

a. Participants were asked how many distinct tattoos they have. Most respondents had fewer than five tattoos; the largest number of tattoos was 53. The average and standard deviation were 4.7 and 6.2, respectively.
b. Participants were asked if they had ever experienced an adverse tattoo reaction: 31 (10.3%)answered that they had.
c. Participants were asked what type of adverse reaction to their tattoos they may have experienced. On average, the 13 participants with acute (temporary) adverse reactions had more tattoos (15.7) than the 18 participants with chronic (long-lasting) reactions (7.1) or the 269 participants who reported no adverse reactions (4.0).
d. Participants were also asked if they had any documented allergies. Among respondents with chronic tattoo reactions, 61.1% (11/18) reported having allergies. In contrast, the allergy rate was 38.5% (5/13) among respondents with acute tattoo reactions, and 32.7% (88/269) among those with no adverse tattoo reaction.

25.9 Spider mating. Male orb-weaver spiders (*Cyrtophora citricola*) are used by the females as a protein source right after mating. Researchers reared male and female orb-weaver spiders in the lab to study their mating behavior in adulthood. Here are some of the research questions examined in their published report.[9] For each objective, determine which inference procedure from Chapters 17 through 24 you would consider and state the corresponding null and alternative hypotheses.

a. The males in the study all attempted to mate, but not all were successful. Given that the males are used as a food source after mating, is there evidence that male orb-weaver spiders that mate successfully are significantly heavier than those that are unsuccessful in their mating attempt?

b. The researchers observed a random sample of mating pairs. The spiders had all been identified and measured before the mating occurred spontaneously. Is there evidence that male and female orb-weaver spiders in mating pairs differ significantly in body length?

COMPREHENSIVE REVIEW EXERCISES: ANALYSIS AND CONCLUSION

These exercises cover the range of statistical inference procedures presented in Chapters 17 through 24. Use the specific technology or tables of critical values taught in your class to obtain P-values and confidence intervals. (In practice, statistical software is used to obtain fast and accurate inference computations.) For exercises that ask you to conduct the entire data analysis, your answers should include the **Plan, Solve,** *and* **Conclude** *steps of the four-step process. It is helpful to also summarize in your own words the* **State** *information given in the exercise.*

25.10 Tattoos: inference. Go back to Exercise 25.8. Write a short paragraph interpreting the study findings in context, citing the results of the relevant inference procedures (provided here in alphabetical order):

ANOVA: F=25.1,P≈0$\text{ANOVA}: F = 25.1, P \approx 0$

Chi-square test: X2=6.09, P=0.0475$\text{Chi-square test}: X^2 = 6.09, P = 0.0475$

95% CI for μ: (4.00, 5.40)$95\%\ \text{CI for}\ \mu: (4.00, 5.40)$

95% CI for p: (0.069, 0.138)$95\%\ \text{CI for}\ p: (0.069, 0.138)$

25.11 Ginkgo extract and the post-lunch dip. The post-lunch dip is the drop in mental alertness after a midday meal. Does an extract of the leaves of the ginkgo tree reduce the post-lunch dip? Assign healthy people aged 18 to 40 years to take either ginkgo extract or a placebo pill. After lunch, ask them to read seven pages of random letters and place an Xx over every e.e. Count the number of misses per line read.[10]

a. What is a placebo and why was one group given a placebo?

b. What is the double-blind method and why should it be used in this experiment?

c. Here are summaries of performance after 13 weeks of either ginkgo extract or placebo:

Group	Group size	Mean	Std. dev.
Ginkgo	21	0.06383	0.01462
Placebo	18	0.05342	0.01549

Is there a significant difference between the two groups? What do these data show about the effect of ginkgo extract?

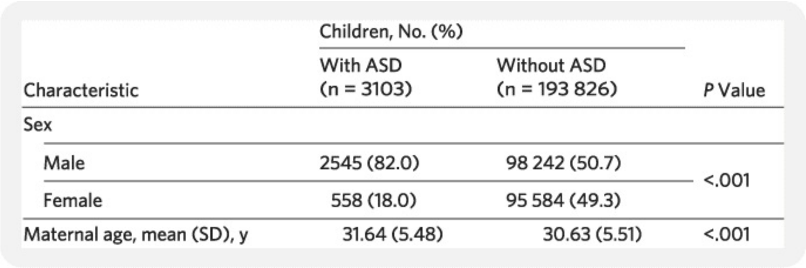
25.12 Autism. A study examined a large cohort of children’s health records from a California health insurance. In this sample, 3103 children had a diagnosis of autism spectrum disorder (ASD) and the remaining 193,826 did not. Studies like this one often provide detailed summary statistics in their published reports, so it is important to be able to read them accurately. Figure 25.3 is a summary table shown as it appears in the publication.[11]

a. Which inference procedure can be used to determine that there is a significant (P<.001)$(P < .001)$ association between sex and ASD diagnosis?

b. Obtain a 95% confidence interval for the proportion of boys in the population of California children with ASD.

c. The publication reports a significant (P<.001)$(P < .001)$ difference in mean maternal age (in years, “y”) depending on the children’s ASD status. Obtain a 95% confidence interval for the difference in mean maternal age of all California children with and without ASD.

d. Write a short paragraph summarizing these findings.



Characteristic	Children, No. (%) With ASD (n = 3103)	Without ASD (n = 193 826)	*P* Value
Sex			
Male	2545 (82.0)	98 242 (50.7)	<.001
Female	558 (18.0)	95 584 (49.3)	
Maternal age, mean (SD), y	31.64 (5.48)	30.63 (5.51)	<.001

Figure 25.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 25.3 Summary statistics provided in a study of autism spectrum disorder (ASD).

25.13 Fish sizes. Table 25.1 contains data on the size of perch caught in a lake in Finland. Examine the relationship between width (y)(y) and length (x)(x) in perch.[12]

a. Plot the data and obtain the equation of the least squares regression line expressing width as a function of length.
b. How well can we predict the width of a perch from its length?
c. The length of a typical perch is about x*=27$x^* = 27$ centimeters. Predict the mean width of such fish and give a 95% confidence interval.
d. Perch number 143 had six newly eaten fish in its stomach. Examine the residuals. Does fish number 143 have an unusually large residual? How might this impact inference?

TABLE 25.1 Measurements on 56 perch

Observation number	Weight (g)	Length (cm)	Width (cm)	Observation number	Weight (g)	Length (cm)	Width (cm)
104	5.9	8.8	1.4	132	197.0	27.0	4.2
105	32.0	14.7	2.0	133	218.0	28.0	4.1
106	40.0	16.0	2.4	134	300.0	28.7	5.1
107	51.5	17.0	2.6	135	260.0	28.9	4.3
108	70.0	18.5	2.9	136	265.0	28.9	4.3
109	100.0	19.2	3.3	137	250.0	28.9	4.6
110	78.0	19.4	3.1	138	250.0	29.4	4.2
111	80.0	20.2	3.1	139	300.0	30.1	4.6
112	85.0	20.8	3.0	140	320.0	31.6	4.8
113	85.0	21.0	2.8	141	514.0	34.0	6.0
114	110.0	22.5	3.6	142	556.0	36.5	6.4
115	115.0	22.5	3.3	143	840.0	37.3	7.8
116	125.0	22.5	3.7	144	685.0	39.0	6.9
117	130.0	22.8	3.5	145	700.0	38.3	6.7
118	120.0	23.5	3.4	146	700.0	39.4	6.3
119	120.0	23.5	3.5	147	690.0	39.3	6.4
120	130.0	23.5	3.5	148	900.0	41.4	7.5
121	135.0	23.5	3.5	149	650.0	41.4	6.0
122	110.0	23.5	4.0	150	820.0	41.3	7.4
123	130.0	24.0	3.6	151	850.0	42.3	7.1

124	150.0	24.0	3.6	152	900.0	42.5	7.2
125	145.0	24.2	3.6	153	1015.0	42.4	7.5
126	150.0	24.5	3.6	154	820.0	42.5	6.6
127	170.0	25.0	3.7	155	1100.0	44.6	6.9
128	225.0	25.5	3.7	156	1000.0	45.2	7.3
129	145.0	25.5	3.8	157	1100.0	45.5	7.4
130	188.0	26.2	4.2	158	1000.0	46.0	8.1
131	180.0	26.5	3.7	159	1000.0	46.6	7.6

25.14 Fish sizes: transforming a variable (optional). We can also use the data in Table 25.1 to study the prediction of the weight of a perch from its length.

a. Describe the pattern of the relationship between weight versus length, with length as the explanatory variable.

b. It is reasonable to expect the one-third power of the weight to have a straight-line relationship with length since animals are three-dimensional. Use your software to create a new variable that is the one-third power of weight. Describe the pattern of this new response variable against length. Is the straight-line pattern now stronger or weaker than that in part a?

c. Find the least-squares regression line to predict the new weight variable from length. Predict the mean of the new variable for perch 27 centimeters long, and give a 95% confidence interval.

25.15 Diabetic mice. The body's natural electrical field helps wounds heal. If diabetes changes this field, that might explain why people with diabetes heal slowly. A study of this idea compared normal mice and mice bred to spontaneously develop diabetes. The investigators attached sensors to the right hip and front feet of the mice and measured the difference in electrical potential (millivolts) between these locations. Here are the data:[13]

Diabetic mice					
14.70	13.60	7.40	1.05	10.55	16.40
10.00	22.60	15.20	19.60	17.25	18.40

9.80	11.70	14.85	14.45	18.25	10.15
10.85	10.30	10.45	8.55	8.85	19.20
Normal mice					
13.80	9.10	4.95	7.70	9.40	8.40
7.20	10.00	14.55	13.30	6.65	8.55
9.50	10.40	7.75	8.70	8.85	12.60

a. Create a dotplot of each sample of electric potentials. There is a low outlier in the diabetic group. Does it appear that potentials in the two groups differ in a systematic way?

b. Is there significant evidence of a difference in mean potentials between the two groups?

c. Repeat your inference without the outlier. Does the outlier affect your conclusion?

25.16 Very-low-birth-weight babies. Progress in medical technology has allowed babies with very low birth weight (VLBW, less than 1500 grams, about 3.3 pounds) to survive without major handicaps at birth, although they may still be vulnerable to long-term issues noticeable in school and adulthood. A long-term study followed 242 VLBW babies to age 20 years, along with a control group of 233 babies from the same population who had normal birth weight.[14]

a. At age 20, 179 of the VLBW group and 193 of the control group had graduated from high school. Is the graduation rate among the VLBW group significantly lower than for the normal-birth-weight controls?

b. IQ scores were available for 113 men in the VLBW group. Their mean IQ score was 87.6, and the standard deviation was 15.1. The 106 men in the control group had a mean IQ score of 94.7, with standard deviation 14.9. Is there good evidence that mean IQ is lower among VLBW men than among controls from similar backgrounds?

c. IQ scores were available for 123 women in the VLBW group and 125 women in the control group. The women in the VLBW group had a mean IQ score of 86.2 (standard deviation 13.4), whereas the women in the control group had a mean IQ score of 89.8 (standard deviation 14.0). Compute a 95% confidence interval for the difference in mean

IQ between VLBW women and normal-birth-weight women from similar backgrounds. What can you conclude from this interval?

d. Can we conclude from this study that being born with a very low birth weight causes lower IQ and lower graduation rates by age 20? Explain your reasoning.

25.17 Healing of skin wounds. Biologists studying the healing of skin wounds measured the rate at which new cells closed a razor cut made in the skin of an anesthetized newt. Here are data from 18 newts, measured in micrometers per hour (μm/h):(μm/h) :[15]

29	27	34	40	22	28	14	35	26
35	12	30	23	18	11	22	23	33

Use a 95% confidence interval to estimate the mean healing rate of razor-cut skin wounds in anesthetized newts. Follow the four-step process.

25.18 Cell-only versus landline users. Random digit dialing telephone surveys do not call cell phone numbers. If the opinions of people who have only cell phones differ from those of people who still have landline service, the poll results may not represent the entire adult population. The Pew Research Center interviewed separate random samples of cell-only and landline telephone users. Here is a two-way table that breaks down these samples by age group:

	Landline sample	**Cell-only sample**
Age 18–29	108	96
Age 30–49	264	70
Age 50–64	202	26
Age 65 or older	178	8
Total	752	200

Is there significant evidence that age distribution differs between landline users and cell phone–only users? Follow the four-step process.

25.19 Does nature heal better? Our bodies have a natural electrical field. In the study presented in Exercise 25.17, the researchers also asked whether a disturbance in this electric field might impede healing. They selected a new random sample of 14 newts. For each newt, the right and left hind limbs were each assigned at random to either experimental or control conditions. The electrical field in the experimental limb was reduced to half its natural value by applying a voltage. The control limb was not manipulated. Table 25.2 gives the rates at which new cells closed a razor cut on the limbs. Is there good evidence that changing the electrical field from its natural level slows healing? Follow the four-step process.

TABLE 25.2 Healing rates (μm/h)$(\mu\mathrm{m}/\mathrm{h})$ when the skin voltage is either reduced (experimental limb) or left unchanged (control limb) in a sample of 14 anesthetized newts

Newt	Experimental limb	Control limb	Newt	Experimental limb	Control limb
1	24	25	8	33	36
2	23	13	9	28	35
3	47	44	10	28	38
4	42	45	11	21	43
5	26	57	12	27	31
6	46	42	13	25	26
7	38	50	14	45	48

25.20 Squirrels and their food supply. That animal species produce more offspring when their supply of food goes up isn't surprising. What is surprising is that some animals appear able to anticipate unusual food abundance. Red squirrels eat seeds from pine cones, a food source that occasionally has very large crops (called seed masting). Here are data on an index of the abundance of pine cones and the average number of offspring per female red squirrel over 16 years:[16]

Cone index xx	0.00	2.02	0.25	3.22	4.68	0.31	3.37	3.09
Offspring yy	1.49	1.10	1.29	2.71	4.07	1.29	3.36	2.41
Cone index xx	2.44	4.81	1.88	0.31	1.61	1.88	0.91	1.04
Offspring yy	1.97	3.41	1.49	2.02	3.34	2.41	2.15	2.12

The strength of the relationship between abundance and offspring is remarkable because the offspring are conceived in the spring, *before* the cones mature in the fall to feed the new young squirrels through the winter. How significant is the evidence that a greater number of cones is associated with more offspring in the same year? Follow the four-step process.

25.21 Soybean genetics. A soybean mutant exhibiting a tan-colored seed coat was recently identified. Crossing the mutant with the wild-type yellow-seeded soybean plant gave all yellow-seeded soybeans in the first generation, F1. Self-pollination of F1 resulted in a second generation, F2, consisting of 814 yellow-seeded and 251 tan-seeded plants. This pattern suggests a simple dominant-recessive model involving a single genetic mutation (3:1 expected ratio; 3/4 and 1/4 expected proportions).[17] Do the data support this interpretation? Run the appropriate chi-square test and conclude in context. Follow the four-step process.

25.22 Comparing tropical flowers Ethan Temeles of Amherst College, with his colleague W. John Kress, studied the relationship between varieties of the tropical flower *Heliconia* on the island of Dominica and the different species of hummingbirds that fertilize the flowers.[18] Over time, the researchers believe, the lengths of the flowers and the form of the hummingbirds' beaks have evolved to match each other. If that is true, flower varieties fertilized by different hummingbird species should have distinct distributions of length. Table 25.3 gives length measurements (in millimeters, mm) for samples of three varieties of *Heliconia,* each fertilized by a different species of hummingbird. Do the three varieties have different flower lengths, on average? Follow the four-step process.

TABLE 25.3 Flower lengths (millimeters) for three *Heliconia* varieties

H. bihai	47.12	46.75	46.81	47.12	46.67	47.43	46.44	46.64
	48.07	48.34	48.15	50.26	50.12	46.34	46.94	48.36
***H. caribaea* red**	41.90	42.01	41.93	43.09	41.47	41.69	39.78	40.57
	39.63	42.18	40.66	37.87	39.16	37.40	38.20	38.07

	38.10	37.97	38.79	38.23	38.87	37.78	38.01	
H. caribaea **yellow**	36.78	37.02	36.52	36.11	36.03	35.45	38.13	37.10
	35.17	36.82	36.66	35.68	36.03	34.57	34.63	

25.23 Temperature response in goldfish. Goldfish are cold-blooded and have strong physiological responses to sudden temperature changes (“acute response”). An experiment compared the acute response to a given test temperature (10, 12, 15, 22, or 25 degrees Celsius) in goldfish who had been kept for months in tanks with water that was 22 degrees Celsius. Table 25.4 gives the ventilation rate in number of opercular beats per minute for each fish in this experiment.[19] Is there evidence that goldfish ventilation rate varies significantly when the fish are exposed suddenly to different temperatures? Follow the four-step process.

TABLE 25.4 Ventilation rates (number of opercular beats per minute) of 90 goldfish kept at different temperatures (degrees Celsius)

Water temperature (°C)				
10	**12**	**15**	**22**	**25**
43	25	64	114	104
39	30	69	86	158
8	68	76	120	160
30	54	82	75	120
44	63	78	40	141
64	35	49	96	79
12	80	96	90	122
15	80	43	65	172
47	71	62	54	110
14	12	48	86	153
28	27	46	87	96
40	8	54	20	110
70	15	30	78	128
21	40	70	80	158
47	30	48	87	117
20	4	66	112	85

34	45	48	67	189
35	30	46	52	120

25.24 Drug interactions. A drug treatment can sometimes be affected by the use of another drug taken concomitantly. This is why doctors and pharmacists routinely ask patients about all the drugs they use, even the ones available over the counter. One study examined the impact of nonsteroidal anti-inflammatory drugs (NSAIDs) on the efficacy of the antidepressant citalopram in clinically depressed individuals. Data from a large clinical trial were reviewed, and clinical remission from depressive symptoms was assessed for individuals who had taken NSAIDs concomitantly and those who had not. Here are the findings:[20]

	Remission	**No remission**
NSAID use	182	227
No NSAID use	628	509

How strong is the evidence that concomitant NSAID use lowers the effectiveness of the antidepressant citalopram? Follow the four-step process.

LARGE DATA SET EXERCISES

These exercises guide you through the comprehensive analysis of more complex data sets characterized by a large number of observations, a number of variables, or both. They cover concepts and approaches presented in Chapters 17 through 24. Note that short answers are not available for these exercises to enable their use for class work or assignments.

25.25 Ant morphology. Researchers were interested in the epigenetic factors that may influence quantitative traits such as size in the carpenter ant (*Camponotus floridanus*).[21] The file *Large.Ant* contains data on the scape length (in microns; a simple and commonly used proxy for body size) of a sample of 179 adult worker carpenter ants. The researchers concluded that the observed variations in body proportions are indicative of two types of workers among adult carpenter ants, which they labeled major workers and minor workers.

a. Describe the distribution of scape lengths among the major workers and obtain a 95% confidence interval for the mean scape length in the population of carpenter ant major workers.

b. Describe the distribution of scape lengths among the major workers and obtain a 95% confidence interval for the mean scape length in the population of carpenter ant minor workers.

c. Estimate with 95% confidence the difference in mean scape length in the populations of major worker and minor worker carpenter ants. Is this difference statistically different?

d. Does your analysis support the researchers’ conjecture of two subpopulations of worker carpenter ants with different body size distributions? Write a short paragraph summarizing your findings.

25.26 Essential tremor. Essential tremor is neurological movement disorder characterized by involuntary rhythmic movement, which typically interferes with the full use of the arms and hands. The most common movement disorder, it differs from Parkinson's disease in that it is typically action related. Researchers recruited a representative sample of 15 subjects diagnosed with essential tremor to test the efficacy of a tremor-canceling device. Before proceeding with the experimental testing, the researchers obtained the baseline characteristics of the participating subjects.[22] The findings, as displayed in the original publication, are available in the *Large.Tremor* data file. "TRS" stands for Tremor Rating Scale, a measure of tremor severity. "DBS" stands for deep brain stimulation, a surgical treatment sometimes used for patients with essential tremor.

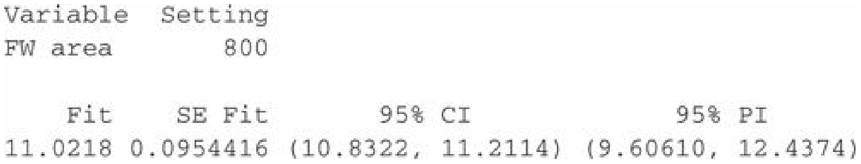
LARGE DATA SET

a. Estimate the mean TRS total score in a population of individuals diagnosed with essential tremor similar to those enrolled in the study.
b. Essential tremor can occur at any age (though it is most common in people older than age 40) and the condition often worsens over time. If appropriate, test whether the relationship between TRS total score and patient age is statistically significant and obtain a 95% confidence interval for the slope of the relationship.
c. If appropriate, test whether the relationship between TRS total score and disease duration is statistically significant and obtain a 95% confidence interval for the slope of the relationship.
d. For which of the relationships in parts b and c, if any, would you be willing to use inference for prediction? Explain your reasoning.

25.27 Butterfly wings. Monarch butterflies (*Danaus plexippus*) undertake large seasonal migrations over North America. Researchers measured the forewing weight (in mg) and surface area (in mm^2) in a sample of 92 monarch butterflies, all of which were reared in captivity in identical conditions.[23] The data are available in the *Large.Butterfly* data file.

LARGE DATA SET

a. The relationship between surface area (explanatory) and weight of the forewing is roughly linear. Verify that the conditions for inference are met and test whether the relationship is statistically significant.
b. Obtain a 95% confidence interval for the slope of the relationship and interpret it in context.
c. Use the following Minitab output to predict with 95% confidence the forewing weight for one monarch butterfly whose forewing has a surface area of 800 mm2.$800\ \text{mm}^2$.



```
Variable  Setting
FW area       800

    Fit    SE Fit       95% CI             95% PI
11.0218 0.0954416 (10.8322, 11.2114) (9.60610, 12.4374)
```

25.28 More on butterfly wings. Continue your analysis of the data in Exercise 25.27 about the migratory butterfly species *Danaus plexippus*. Be sure to check the conditions for each inference procedure you perform.

a. Estimate the mean forewing area of monarch butterflies with 95% confidence.
b. Obtain a 95% confidence interval for the mean forewing weight of monarch butterflies.
c. Is there evidence that male and female monarch butterflies differ in mean forewing surface area?
d. Is there evidence that the mean forewing weight differs in male and female monarch butterflies?
e. Use your work in parts a through d to write a short paragraph describing the morphology of the forewing in monarch butterflies.

25.29 Lung capacity of children. Forced expiratory volume (FEV, measured in liters) is a primary indicator of lung function and corresponds to the volume of air that can forcibly be blown out in the first second after full inspiration. The *Large.FEV* data file contains the FEV values of a large sample of children, along with some categorical descriptors of each individual.[24] Create a new variable based on age: *preschool* for ages 3 to 5, *elementary* for ages 6 to 10, *middle* for ages 11 to 13, and *highschool* for ages 14 and older.

a. Is there evidence that FEV is not the same for all four age groups? Verify that the conditions for ANOVA are met, run the test, and conclude in context.
b. A school district is looking for reference FEV values for children. Obtain a 95% confidence interval for the mean FEV of children in each school-age group (elementary, middle, and high school).
c. For each school-age group, estimate with 95% confidence how much boys and girls in that age group differ in mean FEV. [One issue with this approach is that multiple confidence intervals do not guarantee that *all* of them would capture their respective parameter with an overall high confidence level. Chapter 26 describes a method (two-way ANOVA and follow-up analysis) to handle situations like this one with two explanatory variables.]

25.30 Children's growth. The data file *Large.FEV* described in Exercise 25.29 also includes data on the children's height (in inches). For each of the four age groups (*preschool, elementary, middle, highschool*), assess how much boys and girls differ in mean height by computing a 95% confidence interval for µB-µG.$\mu_B - \mu_G$. Verify the conditions for inference and write a short description of your findings. [One issue with this approach is that multiple confidence intervals do not guarantee that *all* of them would capture their respective parameter with an overall high confidence level. Chapter 26 describes a method (two-way ANOVA and follow-up analysis) to handle situations like this one with two explanatory variables.]

25.31 More on the lung capacity of children. In Exercise 8.20 (page 209), you examined how children's growth affects lung capacity by describing the relationship between FEV and age for all the individuals in the data file *Large.FEV*. The analysis showed that the pattern of association between FEV and age is different for adolescents (middle school and high school: ages 11 and older) and for younger children (preschool and elementary school: ages 10 and younger).

a. Among the younger children, boys and girls are similar enough to be studied together. Find a 95% confidence interval for the slope of the population regression line between FEV and height for these younger children. Verify that the conditions for inference are satisfied. Obtain a 95% confidence interval for the mean FEV of children who are 5 feet tall (60 inches).

b. Study the relationship between FEV and height for adolescent girls. Verify that the conditions for inference are satisfied and obtain a 95% confidence interval for the slope of the population regression line. Obtain a 95% confidence interval for the mean FEV of adolescent girls who are 68 inches tall.

c. Study the relationship between FEV and height for adolescent boys. Verify that the conditions for inference are satisfied and find a 95% confidence interval for the slope of the population regression line. Obtain a 95% confidence interval for the mean FEV of adolescent boys who are 68 inches tall. Compare this interval with the one you obtained in part b.

ONLINE DATA SOURCES

Here are a few regularly updated, online data sources relevant to the life sciences that can help further illustrate and practice the concepts introduced in Chapters 17 through 24.

- datadryad.org

  The Dryad Digital Repository, providing data from a variety of published scientific journals that have been offered free for use.
- www.gallup.com

  Gallup surveys, a number of which focus on health topics
- www.pewresearch.org/science/

  Pew Research Center, section on scientific topics (surveys)
- www.plos.org

  PLOS publications, a set of peer-reviewed journals with a focus on the life sciences. All PLOS articles are “Open Access” (free, unrestricted access and reuse) and some even make their raw data available.
- www.sciencedaily.com

  ScienceDaily: portal for newly published science articles provided with a summary in layperson’s terms and a link to the original source material
- archive.ics.uci.edu/ml

  UCI Machine Learning Repository: a large repository of data sets free for use with proper citation

EESEE CASE STUDIES

The Electronic Encyclopedia of Statistical Examples and Exercises (EESEE) is available on the companion website. These more elaborate stories, with data, provide settings for longer case studies. Here are some suggestions for EESEE stories that apply the ideas you have learned in Part III.

- **Blood Alcohol Content.** All questions.
- **Emissions from an Oil Refinery.** Both questions.
- **Is Caffeine Dependence Real?** Questions 2, 3, 4, and 6.
- **Leave Survey after the Beep.** Question 3.
- **Passive Smoking and Respiratory Health.** Questions 1, 3, and 4.
- **Radar Detectors and Speeding.** Questions 1, 3, and 5.
- **Seasonal Weevil Migration.** Question 1.
- **Stress among Pets and Friends.** All questions.
- **Surgery in a Blanket.** All questions.

NOTES AND DATA SOURCES

“About This Book”

1. B. Baldi and J. Utts, “What your future doctor should know about statistics: Must-include topics for introductory undergraduate biostatistics,” *American Statistician*, 69 (2015), pp. 231–240. doi:10.1080/00031305.2015.1048903.
2. The 2005 GAISE College Report can be found at www.amstat.org/education/gaise/GaiseCollege_full.pdf.
3. GAISE College Report ASA Revision Committee, “Guidelines for Assessment and Instruction in Statistics Education College Report 2016,” www.amstat.org/education/gaise.
4. D. Neumann et al., “real-life data when teaching statistics: Student perceptions of this strategy in an introductory statistics course,” *Statistics Education Research Journal*, 12 (2013), pp. 59–70.
5. R. L. Wasserstein and N. A. Lazar, “The ASA’s statement on p-values: Context, process, and purpose,” *American Statistician*, 70 (2016), pp. 129–133. doi:10.1080/00031305.2016.1154108.
6. The 2016 GAISE report (see Note 2) emphasizes that tables of z,t,X2, z, t, X^2, and FF values do not reflect modern statistical practice. Therefore, learning to use them is no longer a relevant skill outside the classroom.
7. G. Gigerenzer and A. Edwards, “Simple tools for understanding risks: From innumeracy to insight,” *British Medical Journal*, 327 (2003), pp. 741–744.
8. L. D. Brown, T. Cai, and A. DasGupta, “Interval estimation for a binomial proportion,” *Statistical Science*, 16 (2001), pp. 101–133.

“Statistical Thinking”

1. S. Paruthi et al., “Recommended amount of sleep for pediatric populations: A consensus statement of the American Academy of Sleep Medicine,” *Journal of Clinical Sleep Medicine*, 12 (2016), pp. 785–

786. N. F. Watson et al., “Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society,” *Sleep,* 38 (2015), pp. 843–844.

2. E. W. Campion, “Editorial: Power lines, cancer, and fear,” *New England Journal of Medicine,* 337 (1997). The study report is M. S. Linet et al., “Residential exposure to magnetic fields and acute lymphoblastic leukemia in children,” in the same issue. See also G. Taubes, “Magnetic field–cancer link: Will it rest in peace?,” *Science,* 277 (1997), p. 29.

3. L. E. Taylor et al., “Vaccines are not associated with autism: An evidence-based meta-analysis of case-control and cohort studies,” *Vaccine,* 32 (2014), pp. 3623–3629. doi:10.1016/j.vaccine.2014.04.085.

4. K. Hartmann et al., “Outcomes of routine episiotomy: a systematic review,” *Journal of the American Medical Association,* 293 (2005), pp. 2141–2148.

5. “Energy drink consumption and its association with sleep problems among U.S. service members on a combat deployment—Afghanistan, 2010,” *Morbidity and Mortality Weekly Report*, 61 (2012), pp. 895–898.

6. See, for example, M. Enserink, “The vanishing promises of hormone replacement,” *Science,* 297 (2002), pp. 325–326; and B. Vastag, “Hormone replacement therapy falls out of favor with expert committee,” *Journal of the American Medical Association,* 287 (2002), pp. 1923–1924. A National Institutes of Health panel’s comprehensive report is *International Position Paper on Women’s Health and Menopause,* NIH Publication 02-3284, 2002.

7. The data are part of a larger data set in the *Journal of Statistics Education* archive, accessible online. The original source is P. Brofeldt, “Bidrag till kaennedom on fiskbestondet i vaara sjoear: Laengelmaevesi,” in T. H. Jaervi, *Finlands fiskeriet*, Vol. 4, *Meddelanden utgivna av fiskerifoereningen i Finland*, Helsinki, 1917. The data were contributed to the archive (with information in English) by Juha Puranen of the University of Helsinki.

8. We thank Rudi Berkelhamer of the University of California at Irvine for these data. The data are part of a larger set collected for an undergraduate lab exercise in scientific methods.

9. A. C. Nielsen, Jr., “Statistics in marketing,” in *Making Statistics More Effective in Schools of Business*, Graduate School of Business, University of Chicago, 1986.
10. J. S. Mandelblatt et al., “Effects of mammography screening under different screening schedules: Model estimates of potential benefits and harms,” *Annals of Internal Medicine*, 151 (2009), pp. 738–747.

Chapter 1

1. Discrete variables can also be countably infinite, as in the enumeration of all the whole numbers from zero to infinity. We will see examples of such variables in optional Chapter 12, in the context of Poisson distributions.
2. M. Joly et al., “Posture does not matter! Paw usage and grasping paw preference in a small-bodied rooting quadrupedal mammal,” *PLOS ONE*, 7 (2012): e38228. doi:10.1371/journal.pone.0038228.
3. S. B. Willingham et al., “The CD47-signal regulatory protein alpha (SIRPa) interaction is a therapeutic target for human solid tumors,” *Proceedings of the National Academy of Sciences*, 109 (2012), pp. 6662–6667.
4. The *Electronic Encyclopedia of Statistical Examples and Exercises* (EESEE) is a free resource for statistical education. Many of the examples in the EESEE database are available on the text’s companion website, www.whfreeman.com/psls3e.
5. NPR (National Public Radio) News Service article from *All Things Considered*, “‘A concussion can lead to sleep problems that last for years,” April 27, 2016, www.npr.org.
6. Substance Abuse and Mental Health Services Administration, “Behavioral health barometer: United States, 2015.”
7. C. A. Roberto et al., “Influence of licensed characters on children’s taste and snack preferences,” *Pediatrics*, 126 (2010), peds.2009-3433. doi:10.1542/peds.2009-3433.
8. Data provided by Chris Olsen, who found the information in scuba-diving magazines.

9. Our eyes do respond to area, but not quite linearly. It appears that we perceive the ratio of two bars to be about the 0.7 power of the ratio of their actual areas. See W. S. Cleveland, *The Elements of Graphing Data,* Wadsworth, 1985, pp. 278–284.

10. L. J. Paulozzi et al., “Vital signs: Variation among states in prescribing of opioid pain relievers and benzodiazepines – United States, 2012,” *CDC Morbidity and Mortality Weekly Report,* July 4, 2014.

11. California Department of Public Health, “Pertussis report, June 1, 2015.”

12. Centers for Disease Control and Prevention (CDC) website, www.cdc.gov, Lyme disease statistics.

13. Statistical Society of Canada, 2004 case study II. Data contributed by Caroline Davis, Elizabeth Blackmore, Deborah Katzman, and John Fox.

14. D. L. Arsenau, “Comparison of diet management instruction for patients with non-insulin dependent diabetes mellitus: learning activity package vs. group instruction,” MS thesis, Purdue University, 1993.

15. U.S. Geological Survey, National Water Information System, Shark River, Everglades National Park, Florida, waterdata.usgs.gov.

16. NASA Goddard Institute for Space Studies (GISS), “Surface temperature analysis,” http://data.giss.nasa.gov/gistemp.

17. National Institute on Drug Abuse, “Overdose deaths rates,” www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates.

18. National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Trends in Atmospheric Carbon Dioxide, www.esrl.noaa.gov.

19. From the UCI Machine Learning Archive at archive.ics.uci.edu/ml.

20. D. S. Ludwig et al., “Pregnancy weight gain and childhood body weight: A within-family comparison,” *PLOS Med,* 10:e1001521 (2013). doi:10.1371/journal.pmed.1001521.

21. S. Keats and S. Wiggins,“Future diets: Implications for agriculture and food prices.” *Odi*, (2014), retrieved from www.odi.org.uk.

22. G. P. Sutton and M. Burrows, “Biomechanics of jumping in the flea,” *Journal of Experimental Biology,* 214 (2011), pp. 836–847. doi:10.1242/jeb.052399.

23. See Note 4.

24. National Center for Health Statistics, “Deaths and death rates for the 10 leading causes of death in specified age groups: United States, preliminary 2010,” *National Vital Statistics Reports,* 60 (2012), Table 7.

25. Florida Fish and Wildlife Conservation Commission, Marine Mammal Pathobiology Laboratory, Final Manatee Mortality Table, 2012.

26. A. K. R. LePort et al., “Behavioral and neuroanatomical investigation of Highly Superior Autobiographical Memory (HSAM),” *Neurobiology of Learning and Memory,* 98 (2012), pp. 78–92.

27. This data set was provided by Nicolas Fisher.

28. National Institutes of Health, Essential Fatty Acids Education site, efaeducation.nih.gov.

29. Data provided by Drina Iglesia, Purdue University. These data are part of a larger study reported in D. D. S. Iglesia, E. J. Cragoe, Jr., and J. W. Vanable, “Electric field strength and epithelization in the newt (*Notophthalmus viridescens*),” *Journal of Experimental Zoology,* 274 (1996), pp. 56–62.

30. Nutrition, Physical Activity and Obesity Data, Trends and Maps web site (Atlanta, GA: U.S. Department of Health and Human Services, CDC, National Center for Chronic Disease Prevention and Health Promotion, Division of Nutrition, Physical Activity and Obesity, 2015), www.cdc.gov/nccdphp/DNPAO.

31. We thank Ethan J. Temeles of Amherst College for providing these data. His work is described in E. J. Temeles and W. J. Kress, “Adaptation in a plant–hummingbird association,” *Science*, 300 (2003), pp. 630–633.

32. National Oceanic and Atmospheric Administration, www.beringclimate.noaa.gov.

33. Y.-H. Huang et al., “Nanoparticle-delivered suicide gene therapy effectively reduces ovarian tumor burden in mice,” *Cancer Research,* 69 (2009), pp. 6184–6191.

34. I. Heap, “International survey of herbicide resistant weeds,” www.weedscience.org.

35. www.gallup.com/poll/184403/far-2015-americans-exercising-frequently.aspx.

36. See Note 15.

Chapter 2

1. Data collected by Brigitte Baldi.

2. This isn’t a mathematical theorem. The mean can be less than the median in right-skewed distributions that include only a few values, many of which lie exactly at the median. The rule almost never fails for distributions that include a large number of values, and counterexamples don’t appear clearly skewed in graphs even though they may be slightly skewed according to technical measures of skewness. See P. T. von Hippel, “Mean, median, and skew: Correcting a textbook rule,” *Journal of Statistics Education*, 13 (2005), www.amstat.org/publications/jse.

3. E. Spanger-Siegfried et al., “Encroaching tides: How sea level rise and tidal flooding threaten U.S. East and Gulf Coast communities over the next 30 years.” Cambridge, MA: Union of Concerned Scientists, (2014), www.ucsusa.org/encroachingtides.

4. J. Lanett Beltran, “The effect of spider size and fibroin composition on the material properties of dragline silk from *Nephila clavipes*,” MS thesis, University of California at Irvine, 2007.

5. A. Dell’Anno and R. Danovaro, “Extracellular DNA plays a key role in deep-sea ecosystem functioning,” *Science*, 309 (2005), pp. 2179.

6. T. Yamada et al., “A new computer-based pediatric vision-screening test,” *Journal of AAPOS*, 19 (2015), pp. 157–162. doi:10.1016/j.jaapos.2015.01.011.

7. Although this is not mathematically correct, you might find it helpful to think of the standard deviation s as representing the average space (or absolute distance) between the n data points and their mean.

8. J. Golding et al., “Dietary predictors of maternal prenatal blood mercury levels in the ALSPAC birth cohort study,” *Environmental Health Perspectives*, 121 (2013), pp. 1214–1218.

doi:10.1289/ehp.1206115.

9. R. Y. Hsia et al., “Variation in charges for 10 common blood tests in California hospitals: A cross-sectional analysis.” *BMJ Open* 4 (2014), e005482–e005482. doi:10.1136/bmjopen-2014-005482.

10. EESEE story “Acorn size and oak tree range.”

11. S. Monk, “10 factors that can affect blood pressure readings,” January 2010, SunTech Medical, www.suntechmed.com.

12. See Note 26 for Chapter 1.

13. D. J. Madigana et al., “Pacific bluefin tuna transport Fukushima-derived radionuclides from Japan to California,” *Proceedings of the National Academy of Sciences*, 109 (2012), pp. 9483–9486. doi:10.1073/pnas.1204859109.

14. G. Dezecache and R. I. M. Dunbar, “Sharing the joke: the size of natural laughter groups,” *Evolution and Human Behavior*, 33 (2012), pp. 775–779, doi:10.1016/j.evolhumbehav.2012.07.002.

15. T. Bjerkedal, “Acquisition of resistance in guinea pigs infected with different doses of virulent tubercle bacilli,” *American Journal of Hygiene*, 72 (1960), pp. 130–148.

16. M. A. Laskey et al., “Bone changes after 3 mo of lactation: influence of calcium intake, breast-milk output, and vitamin D–receptor genotype,” *American Journal of Clinical Nutrition*, 67 (1998), pp. 685–692.

17. L. H. Yang, “Periodical cicadas as resource pulses in North American forests,” *Science*, 306 (2004), pp. 1565–1567. The data are simulated Normal values that match the means and standard deviations reported in this article.

18. J. A. Levine et al., “Inter-individual variation in posture allocation: possible role in human obesity,” *Science*, 307 (2005), pp. 584–586. We thank James Levine for providing the data.

19. We thank Charles Cannon of Duke University for providing the data. The study report is C. H. Cannon, D. R. Peart, and M. Leighton, “Tree species diversity in commercially logged Bornean rainforest,” *Science*, 281 (1998), pp. 1366–1367.

20. J. Tunga et al., “Social environment is associated with gene regulatory

variation in the rhesus macaque immune system," *Proceedings of the National Academy of Sciences,* 109 (2012), pp. 6490–6495, doi:10.1073/pnas.1202734109.

21. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by John P. Holcomb, Jr., of Cleveland State University.

Chapter 3

1. From a graph in N. E. Cantin et al., "Ocean warming slows coral growth in the Central Red Sea," *Science,* 329 (2010), pp. 322–325. doi:10.1126/science.1190182.
2. Pew Research Center, "Global concern about climate change, broad support for limiting emissions," November 2015.
3. Florida Fish and Wildlife Conservation Commission, "Boating accident statistics" and "Manatee mortality," www.myfwc.com.
4. A careful study of this phenomenon is W. S. Cleveland, P. Diaconis, and R. McGill, "Variables on scatterplots look more highly correlated when the scales are increased," *Science,* 216 (1982), pp. 1138–1141.
5. From a graph in J. A. Levine, N. L. Eberhardt, and M. D. Jensen, "Role of nonexercise activity thermogenesis in resistance to fat gain in humans," *Science,* 283 (1999), pp. 212–214.
6. M. E. Peterson et al., "The dependence of enzyme activity on temperature: determination and validation of parameters," *Biochemical Journal,* 402 (2007), pp. 331–337. doi:10.1042/BJ20061143.
7. C. N. Templeton, E. Greene, and K. Davis, "Allometry of alarm calls: black-capped chickadees encode information about predator size," *Science,* 308 (2005), pp. 1934–1937.
8. From a graph in L. Partridge and M. Farquhar, "Sexual activity reduces lifespan of male fruitflies," *Nature,* 294 (1981), pp. 580–582.
9. M. Tamaki et al., "Night watch in one brain hemisphere during sleep associated with the first-night effect in humans," *Current Biology,* 26 (2016), pp. 1190–94. doi:10.1016/j.cub.2016.02.063.
10. P. Eastwick et al., "Mispredicting distress following romantic breakup: Revealing the time course of the affective forecasting error," *Journal of Experimental Social Psychology,* 44 (2008), pp. 800–807.

doi:10.1016/j.jesp.2007.07.001.

11. B. E. Carlson et al., “Meek males and fighting females: Sexually-dimorphic antipredator behavior and locomotor performance is explained by morphology in bark scorpions (*Centruroides Vittatus*),” *PLOS ONE*, 9 (2014), e97648. doi:10.1371/journal.pone.0097648.

12. These data were originally collected by L. M. Linde of UCLA but were first published by M. R. Mickey, O. J. Dunn, and V. Clark, “Note on the use of stepwise regression in detecting outliers,” *Computers and Biomedical Research*, 1 (1967), pp. 105–111. The data have been used by several authors. These are taken from N. R. Draper and J. A. John, “Influential observations and outliers in regression,” *Technometrics*, 23 (1981), pp. 21–26.

13. From graphs in K. Lebret et al., “Grazing resistance allows bloom formation and may explain invasion success of *Gonyostomum semen*,” *Limnology and Oceanography*, 57 (2012), pp. 727–734. doi:10.4319/lo.2012.57.3.0727.

14. C. J. Lawrence, “Post-learning mechanisms related to memory for emotional stimuli in humans,” MS thesis, University of California at Irvine, 2007.

15. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by G. Andy Chang of Youngstown State University.

16. M. Mozny et al., “The impact of climate change on the yield and quality of Saaz hops in the Czech Republic,” *Agricultural and Forest Meteorology*, 149 (2009), pp. 913–919.

17. J. M. Padowski et al., “Neurochemical correlates of caudate atrophy in Huntington’s Disease,” *Movement Disorders*, 29 (2014), pp. 327–335. doi:10.1002/mds.25801.

18. R. Margaria et al., “Energy cost of running,” *Journal of Applied Physiology*, 18 (1963), pp. 367–370.

19. S. J. Husson et al., “Optogenetic analysis of a nociceptor neuron and network reveals ion channels acting downstream of primary sensors,” *Current Biology* 22 (2012), pp. 743–752. doi:10.1016/j.cub.2012.02.066.

20. Data from many studies compiled in D. F. Greene and E. A. Johnson, “Estimating the mean annual seed production of trees,” *Ecology*, 75 (1994), pp. 642–647.

21. C. Tomasetti and B. Vogelstein, “Variation in cancer risk among tissues can be explained by the number of stem cell divisions,” *Science,* 347 (2015), pp. 78–81. doi:10.1126/science.1260825.
22. From a graph in C. G. Wiklund, “Food as a mechanism of density-dependent regulation of breeding numbers in the merlin *Falco columbarius,*” *Ecology*, 82 (2001), pp. 860–867.
23. J. S. Brashares et al., “Bushmeat hunting, wildlife declines, and fish supply in West Africa,” *Science,* 306 (2004), pp. 1180–1183. The data used here are found in the online supplementary material. The published analysis omits data for 1999, an extreme low outlier, without explanation.
24. From a graph in N. I. Eisenberger, M. D. Lieberman, and K. D. Williams, “Does rejection hurt? An fMRI study of social exclusion,” *Science,* 302 (2003), pp. 290–292.
25. From a graph in G. J. Tattersall et al., “Heat exchange from the toucan bill reveals a controllable vascular thermal radiator,” *Science,* 325 (2009), pp. 468–470.

Chapter 4

1. University of Oklahoma Police Department, “The Police Notebook,” based on material from the National Highway Traffic Safety Administration, Department of Motor Vehicles, “California Driver Handbook.” The exact definition varies depending on the alcohol content of the particular brand consumed.
2. These terms are typically abbreviated as SST, or Sum of Squares Total, for the denominator. The variation in y^$\hat{y}$ as xx pulls it along the line is noted as SSR, or Sum of Squares due to Regression; the variation of the residuals is noted as SSE, or Sum of Squared Errors. If you plot these vertical deviations on a scatterplot, you can see by simple geometry that SST = SSR + SSE.
3. K. McCrickerd et al., “Subtle changes in the flavour and texture of a drink enhance expectations of satiety,” *Flavour,* (2012), pp. 1–20. doi:10.1186/2044-7248-1-20.

4. From a graph in T. Singer et al., “Empathy for pain involves the affective but not sensory components of pain,” *Science,* 303 (2004), pp. 1157–1162. Data for other brain regions showed a stronger correlation and no outliers.

5. Data from G. A. Sacher and E. F. Staffelt, “Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth,” *American Naturalist,* 108 (1974), pp. 593–613. These data were found in F. L. Ramsey and D. W. Schafer, *The statistical sleuth: A course in methods of data analysis* (Duxbury, 1997).

6. T. Stockwell et al., “Do ‘moderate’ drinkers have reduced mortality risk? A systematic review and meta-analysis of alcohol consumption and all-cause mortality,” *Journal of Studies on Alcohol and Drugs,* 77 (2016), pp. 185–198. doi:10.15288/jsad.2016.77.185.

7. D. P. Casey et al., “Relationship between muscle sympathetic nerve activity and aortic wave reflection characteristics in young men and women,” *Hypertension,* 57 (2011), pp. 421–427.

8. U.S. Food and Drug Administration (FDA), “FDA requiring lower recommended dose for certain sleep drugs containing zolpidem,” January 10, 2013, www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm33479

9. F. H. Messerli, “Chocolate consumption, cognitive function, and Nobel laureates,” *New England Journal of Medicine,* 367 (2012), pp. 1562–1564. doi:10.1056/NEJMon1211064.

10. L. L. Calderon et al., “Risk factors for obesity in Mexican-American girls: dietary factors, anthropometric factors, physical activity, and hours of television viewing,” *Journal of the American Dietetic Association*, 96 (1996), pp. 1177–1179.

11. *The health consequences of smoking—50 years of progress: a report of the Surgeon General*, U.S. Department of Health and Human Services (2014).

12. FDA, “CDC concludes Zika causes microcephaly and other birth defects,” April 13, 2016, www.cdc.gov/media/releases/2016/s0413-zika-microcephaly.html.

13. P. K. Mills et al., “Dietary habits and breast cancer incidence among

Seventh-Day Adventists,” *Cancer*, 64 (1989), pp. 582–590.

14. J. D. McLister, “The energetics of male reproductive behavior in the treefrog, *Hyla versicolor*,” MS thesis, University of California at Irvine, 2000.

15. S. Yoshida et al., “A bacterium that degrades and assimilates poly(ethylene terephthalate),” *Science*, 351 (2016), pp. 1196–1199. doi:10.1126/science.aad6359.

16. From a presentation by Charles Knauf, Monroe County (New York) Environmental Health Laboratory.

17. Statistical Society of Canada, 2001 Case Study II. Original publication: F. He and R. Duncan, “Density-dependent effects on tree survival in an old-growth Douglas fir forest,” *Journal of Ecology*, 88 (2000), pp. 676–688.

18. From a graph in G. D. Martinsen, E. M. Driebe, and T. G. Whitham, “Indirect interactions mediated by changing plant chemistry: beaver browsing benefits beetles,” *Ecology*, 79 (1998), pp. 192–200.

19. Frank J. Anscombe, “Graphs in statistical analysis,” *American Statistician*, 27 (1973), pp. 17–21.

20. National Toxicology Program, CAS No. 22839-47-0 (2005).

21. From a graph in Feng Sheng Hu et al., “Cyclic variation and solar forcing of Holocene climate in the Alaskan subarctic,” *Science*, 301 (2003), pp. 1890–1893.

22. From a graph in A. L. Perry et al., “Climate change and distribution shifts in marine fishes,” *Science*, 308 (2005), pp. 1912–1915. The explanatory variable is the five-year running mean of winter (December to March) sea-bottom temperature.

23. C. Carbone and J. L. Gittleman, “A common rule for the scaling of carnivore density,” *Science*, 295 (2002), pp. 2273–2276.

24. M. Ding et al., “Association of coffee consumption with total and cause-specific mortality in 3 large prospective cohorts,” *Circulation*, 132 (2015), pp. 2305–2315. doi:10.1161/CIRCULATIONAHA.115.017341.

25. From FiveThirtyEight, “Two times Mike Pence brushed off science,” July 15, 2016, fivethirtyeight.com/features/two-times-mike-pence-brushed-off-science.

26. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by M. J. Kahn of Wheaton College.

Chapter 5

1. A. B. Dufour et al., “Foot pain: Is current or past shoewear a factor? The Framingham Foot Study,” *Arthritis Care and Research*, 61 (2009), pp. 1352–1358. doi:10.1002/art.24733.
2. G. Du Toit et al., “Randomized trial of peanut consumption in infants at risk for peanut allergy,” *New England Journal of Medicine,* 372 (2015), pp. 803–813. doi:10.1056/NEJMoa1414850.
3. J. L. Hoogland, “Why do Gunnison’s prairie dogs give anti-predator calls?,” *Animal Behavior*, 51 (1996), pp. 871–880.
4. L. R. Curtin and R. J. Klein, “Direct standardization (age-adjusted death rates),” *National Center for Health Statistics*, 6 (1995), pp. 1–10.
5. Data set presented by the New York State Department of Health, www.health.ny.gov/diseases/chronic/ageadj.htm.
6. C. R. Charig et al., “Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy,” *British Medical Journal*, 292 (1986), pp. 879–882.
7. Condensed from D. R. Appleton, J. M. French, and M. P. J. Vanderpump, “Ignoring a covariate: An example of Simpson’s paradox,” *American Statistician*, 50 (1996), pp. 340–341.
8. R. B. Turner et al., “An evaluation of *Echinacea angustifolia* in experimental rhinovirus infections,” *New England Journal of Medicine,* 353 (2005), pp. 341–348.
9. “Americans’ coffee consumption is steady, few want to cut back,” Gallup, July 29, 2015.
10. Information for physicians at www.botoxseveresweating.com.
11. G. D. Myer et al., “Youth versus adult weightlifting injuries presenting to United States emergency rooms: accidental versus nonaccidental injury mechanisms,” *Journal of Strength and Conditioning Research,*

23 (2009), pp. 2054–2060.

12. Feeding Infants and Toddlers Study, a Gerber initiative (2002), http://medical.gerber.com/nestlescience/fits.aspx.

13. M. B. Esser et al., “Prevalence of alcohol dependence among US adult drinkers, 2009–2011,” *Preventing Chronic Disease*, 11 (2014), pp. E206. doi:10.5888/pcd11.140329.

14. A. E. Czeizel and I. Dudas, “Prevention of the first occurrence of neural-tube defects by periconceptional vitamin supplementation,” *New England Journal of Medicine*, 327 (1992), pp. 1832–1835.

15. M. Lebwohl et al., “Phase 3 studies comparing brodalumab with ustekinumab in psoriasis,” *New England Journal of Medicine*, 373 (2015), pp. 1318–1328. doi:10.1056/NEJMoa1503824.

16. K. A. Rose et al., “Myopia, lifestyle, and schooling in students of Chinese ethnicity in Singapore and Sydney,” *Archives of Ophthalmology*, 126 (2008), pp. 527. doi:10.1001/archopht.126.4.527.

17. W. Gatling, M. A. Mullee, and RD Hill, “The general characteristics of a community based population,” *Practical Diabetes*, 5 (1989), pp. 104–107.

18. R. Shine, T. R. L. Madsen, M. J. Elphick, and P. S. Harlow, “The influence of nest temperatures and maternal brooding on hatchling phenotypes in water pythons,” *Ecology*, 78 (1997), pp. 1713–1721.

19. J. E. Williams et al., “Anger proneness predicts coronary heart disease risk,” *Circulation*, 101 (2000), pp. 2034–2039.

20. See Note 17 for Chapter 4.

21. N. J. O. Birkmeyer, “Hospital complication rates with bariatric surgery in Michigan,” *Journal of the American Medical Association*, 304 (2010), pp. 435–442.

22. D. Gonzales et al., “Varenicline, an α4β2$\alpha4\beta2$ nicotinic acetylcholine receptor partial agonist, vs sustained-release bupropion and placebo for smoking cessation: A randomized controlled trial,” *Journal of the American Medical Association*, 296 (2006), pp. 47–55. doi:10.1001/jama.296.1.56.

Chapter 6

1. See, for example, M. Enserink, “The vanishing promises of hormone replacement,” *Science*, 297 (2002), pp. 325–326; B. Vastag, “Hormone replacement therapy falls out of favor with expert committee,” *Journal of the American Medical Association*, 287 (2002), pp. 1923–1924; and J. E. Manson et al., “Estrogen therapy and coronary-artery calcification, *New England Journal of Medicine*, 356 (2007), pp. 2591–2602.

2. J. C. Barefoot et al., “Alcoholic beverage preference, diet, and health habits in the UNC Alumni Heart Study,” *American Journal of Clinical Nutrition*, 76 (2002), pp. 466–472.

3. “In U.S., support up for doctor-assisted suicide,” Gallup, May 27, 2015.

4. T. Chang et al., “Inadequate hydration, BMI, and obesity among US adults: NHANES 2009–2012,” *Annals of Family Medicine*, 14 (2016), pp. 320–24. doi:10.1370/afm.1951.

5. See Note 2 for Chapter 1.

6. See Note 22 for Chapter 1.

7. See Note 3 for Chapter 3.

8. According to the American Association for Public Opinon Research, “surveys based on self-selected volunteers do not have [a] known relationship to the target population and are subject to unknown, non-measurable biases.” Yet opt-in polls are unfortunately quite common.

9. D. A. Frederick and B. N. Jenkins, “Height and body mass on the mating market: Associations with number of sex partners and extra-pair sex among heterosexual men and women aged 18–65,” *Evolutionary Psychology*, 13 (2015), pp. 1–14. doi:10.1177/1474704915604563.

10. L. Clark et al., “Efficacy, effort, and cost comparisons of trapping and acetaminophen-baiting for control of brown treesnakes on Guam,” *Human–Wildlife Interactions*, 6 (2012), pp. 222–236.

11. J. P. D. Guidry et al., “On pins and needles: How vaccines are portrayed on Pinterest,” *Vaccine*, 33 (2015), pp. 5051–5056. doi:10.1016/j.vaccine.2015.08.064.

12. R. C. Parker and P. A. Glass, “Preliminary results of double-sample forest inventory of pine and mixed stands with high- and low-density LiDAR," in Kristina F. Connoe (ed.), *Proceedings of the 12th Biennial Southern Silvicultural Research Conference*, U.S. Department of Agriculture, Forest Service, Southern Research Station, 2004.

13. "Legal and ethical issues in RDD cell phone surveys," www.aapor.org.

14. S. J. Blumberg and J. V. Luke, "Wireless substitution: Early release of estimates from the National Health Interview Survey, January–June 2016," National Center for Health Statistics, December 2016, www.cdc.gov/nchs/nhis.htm.

15. For information on the American Community Survey of households (there is a separate sample of group quarters), go to www.census.gov/acs.

16. "2015 National Health Interview Survey (NHIS) survey description," CDC, June 2016.

17. For information on the University of Chicago's General Social Survey, go to http://www3.norc.org/gss+website.

18. D. Dutwin and P. Lavrakas, "Trends in telephone outcomes, 2008–2015," *Survey Practice*, 9 (2016).

19. For a more in-depth analysis of the issues, see these two reports by the Pew Research Center: "Assessing the representativeness of public opinion surveys" (2012) and "The challenges of polling when fewer people are available to be polled" (2015).

20. "Bad samples," www.aapor.org.

21. R. M. Merrill and J. S. Richardson, "Validity of self-reported height, weight, and body mass index: Findings from the National Health and Nutrition Examination Survey, 2001–2006," *Preventing Chronic Disease*, 6 (2009), A121.

22. For more detail on the limits of memory in surveys, see N. M. Bradburn, L. J. Rips, and S. K. Shevell, "Answering autobiographical questions: the impact of memory and inference on surveys," *Science*, 236 (1987), pp. 157–161.

23. The responses on welfare are from a *New York Times*/CBS News Poll reported in *The New York Times*, July 5, 1992. Many other examples appear in T. W. Smith, "That which we call welfare by any other name would smell sweeter," *Public Opinion Quarterly*, 51 (1987), pp. 75–83.

24. "Health reform and the decline of physician private practice," Physician Foundation, 2010. The cover page states "Includes results of Physicians and Health Reform, a survey of 100,000 physicians." The full report can be found at www.physiciansfoundation.org.

25. L. Misegades et al., "Association of childhood pertussis with receipt of 5 doses of pertussis vaccine by time since last vaccine dose, California, 2010," *Journal of the American Medical Association*, 308 (2012), pp. 2126. doi:10.1001/jama.2012.14939.

26. Most cohort studies are prospective. However, some cohort studies select a group of individuals currently enrolled in a health care plan and then search the subjects' past medical records for factors that may explain their current health conditions: Such cohorts are retrospective rather than prospective.

27. L. Trasande et al., "Infant antibiotic exposures and early-life body mass," *International Journal of Obesity*, 37 (2013), pp. 16–23. doi:10.1038/ijo.2012.132.

28. F. X. Bosch et al., "The causal relation between human papillomavirus and cervical cancer," *Journal of Clinical Pathology*, 55 (2002), pp. 244–265. doi:10.1136/jcp.55.4.244.

29. S. Berkowitz et al., "Material need insecurities, control of diabetes mellitus, and use of health care resources," *JAMA Internal Medicine*, 175 (2015), pp. 257–265. doi:10.1001/jamainternmed.2014.6888.

30. P. Krakowiak et al., "Maternal metabolic conditions and risk for autism and other neurodevelopmental disorders," *Pediatrics*, 129 (2012), peds.2011-2583. doi:10.1542/peds.2011-2583.

31. C. G. Bacon et al., "A prospective study of risk factors for erectile dysfunction," *Journal of Urology*, 176 (2006), pp. 217–221.

32. B. Wilt, "How university students sleep," *Jawbone Blog*, April 20, 2016, jawbone.com/blog/university-students-sleep.

33. L. E. Moses and F. Mosteller, "Safety of anesthetics," in J. M. Tanur et al. (eds.), *Statistics: A Guide to the Unknown*, 3rd ed., Wadsworth, 1989, pp. 15–24.

34. J. P. Robinson-Cimpian, “Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions,” *Educational Researcher,* 43 (2014), pp. 171–185. doi:10.3102/0013189X14534297.

35. Jane Goodall Institute, www.janegoodall.org/.

36. M. A. Parada et al., “The validity of self-reported seatbelt use: Hispanic and non-Hispanic drivers in El Paso,” *Accident Analysis and Prevention*, 33 (2001), pp. 139–143.

37. C. Rampell, “You sleep more than you claim you do,” *New York Times,* January 16, 2014.

38. See Note 16 for Chapter 2.

39. M. Lucas et al., “Coffee, caffeine, and risk of depression among women,” *Archives of Internal Medicine,* 171 (2011), pp. 1571–1578.

40. E. Courchesne et al., “Unusual brain growth patterns in early life in patients with autistic disorder: an MRI study,” *Neurology*, 57 (2001), pp. 245–254.

41. D. R. Young et al., “Effects of physical activity and sedentary time on the risk of heart failure,” *Circulation: Heart Failure,* 7 (2014), pp. 21–27. doi:10.1161/CIRCHEARTFAILURE.113.000529.

42. P.-Y. Scarabin, E. Oger, and G. Plu-Bureau, “Differential association of oral and transdermal oestrogen replacement therapy with venous thromboembolism risk,” *Lancet,* 362 (2003), pp. 428–432.

43. H.-C. Yeh et al.,“Smoking, smoking cessation, and risk for type 2 diabetes mellitus: a cohort study,” *Annals of Internal Medicine,* 152 (2010), pp. 10–17.

Chapter 7

1. K. A. Lee et al., “A randomized trial comparing the cost-effectiveness of 2 approaches for treating unilateral nasolacrimal duct obstruction,” *Archives of Ophthalmology*, 130 (2012), pp. 1525–1533. doi:10.1001/archophthalmol.2012.2853.

2. Y. Higuchia et al., “Day light quality affects the night-break response in the short-day plant chrysanthemum, suggesting differential phytochrome-mediated regulation of flowering,” *Journal of Plant Physiology*, 169 (2012), pp. 1789–1796.

doi:10.1016/j.jplph.2012.07.003.

3. L. L. Miao, “Gastric freezing: An example of the evaluation of medical therapy by randomized clinical trials,” in J. P. Bunker, B. A. Barnes, and F. Mosteller (eds.), *Costs, Risks, and Benefits of Surgery,* Oxford University Press, 1977, pp. 198–211.
4. F. Eippert et al., “Direct evidence for spinal cord involvement in placebo analgesia,” *Science,* 326 (2009), p. 404.
5. K. Linde et al., “Acupuncture for patients with migraine: A randomized controlled trial,” *Journal of the American Medical Association*, 293 (2005), pp. 2118–2125.
6. “FTC complaint charges deceptive advertising by POM Wonderful,” FTC, September 27, 2010.
7. P. E. O’Brien et al., “Laparascopic adjustable gastric banding in severely obese adolescents,” *Journal of the American Medical Association,* 303 (2010), pp. 519–526.
8. K. B. Suttle, M. A. Thomsen, and M. E. Power, “Species interactions reverse grassland responses to changing climate,” *Science,* 315 (2007), pp. 640–642. We thank Kenwyn Blakeslee Suttle for providing these data.
9. J. S. Halterman et al., “Randomized controlled trial to improve care for urban children with asthma,” *Archives of Pediatrics and Adolescent Medicine*, 165 (2011), pp. 262–268.
10. A. Pathak et al., “A noninvasive handheld assistive device to accommodate essential tremor: A pilot study,” *Movement Disorders,* 29 (2014), pp. 838–842. doi:10.1002/mds.25796.
11. D. T. Kirkendall, “Creatinine, carbs, and fluids: How important in soccer nutrition?” *Sports Science Exchange,* 94 (2004), pp. 1–6.
12. See Note 15 for Chapter 5.
13. See Note 2 for Chapter 5.
14. A.-M. Chang et al., “Evening use of light-emitting eReaders negatively affects sleep, circadian timing, and next-morning alertness,” *Proceedings of the National Academy of Sciences*, 112 (2015), pp. 1232–1237. doi:10.1073/pnas.1418490112.
15. L.-P. Wang et al., “Efficacy of acupuncture for migraine prophylaxis: a

single-blinded, double-dummy, randomized controlled trial," *Pain*, 152 (2011), pp. 1864–1871.

16. R. E. Sorge et al., "Olfactory exposure to males, including men, causes stress and related analgesia in rodents," *Nature Methods,* 11 (2014), pp. 629–632. doi:10.1038/nmeth.2935.

17. U.S. Department of Health and Human Services, Public Health Service, National Toxicology Program, *Report on Carcinogens, 11th ed.*, Table B: "Agents, substances, mixtures, or exposure circumstances delisted from the *Report on Carcinogens,*" ntp.niehs.nih.gov/ntp/roc/toc11.html.

18. "Birth control methods," WomensHealth.gov.

19. See Note 8 for Chapter 4.

20. J. A. Clayton, "Studying both sexes: A guiding principle for biomedicine," *FASEB Journal,* 30 (2016), pp. 519–524. doi:10.1096/fj.15-279554.

21. W. Zeng et al., "Sympathetic neuro-adipose connections mediate leptin-driven lipolysis," *Cell*, 163 (2015), pp. 84–94. doi:10.1016/j.cell.2015.08.055.

22. L. Loffredo et al., "Dark chocolate acutely improves walking autonomy in patients with peripheral artery disease," *Journal of the American Heart Association,* 3 (2014), e001072. doi:10.1161/JAHA.114.001072.

23. A. S. Attwood et al., "Glass shape influences consumption rate for alcoholic beverages," *PLOS ONE,* 7 (2012), e43007. doi:10.1371/journal.pone.0043007.

24. The difficulties of interpreting guidelines for informed consent and for the work of institutional review boards in medical research are a main theme of Beverly Woodward, "Challenges to human subject protections in U.S. medical research," *Journal of the American Medical Association,* 282 (1999), pp. 1947–1952. The references in this paper point to other discussions. Updated regulations and guidelines appear on the website for the Office for Human Research Protections of the Department of Health and Human Services, www.hhs.gov/ohrp.

25. NPR News Service article from *All Things Considered,* November 16,

2005, "'My Lobotomy': Howard Dully's journey," www.npr.org.

26. K. Hartmann et al., "Outcomes of routine episiotomy: A systematic review," *Journal of the American Medical Association,* 293 (2005), pp. 2141–2148.

27. U.S. Department of Health and Human Services, "Examining oversight of the privacy & security of health data collected by entities not regulated by HIPAA," 2016, www.healthit.gov.

28. C. Duhigg, "How companies learn your secrets," *New York Times,* February 16, 2012.

29. "President discusses stem cell research," August 9, 2001, www.whitehouse.gov.

30. Proposition 71, "Stem Cell Research; Funding; Bonds; Initiative Constitutional Amendment and Statute." Proposition 71 was passed by California voters in November 2004 and resulted in the creation of the California Institute for Regenerative Medicine.

31. "Removing barriers to responsible scientific research involving human stem cells," March 9, 2009, www.whitehouse.gov.

32. J. Kaiser, "NIH moves to lift moratorium on animal–human chimera research," *Science,* August 4, 2016. doi:10.1126/science.aag0732.

33. K. Gourlay, "Tiny opioid patients need help easing into life," *NPR,* March 25, 2016.

34. A. D. Kramer et al., "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences,* 11 (2014), pp. 8788–8790. doi:10.1073/pnas.1320040111.

35. P. C. Tan et al., "Dengue infection and miscarriage: A prospective case control study," *PLOS Neglected Tropical Diseases,* 6 (2012), e1637. doi:10.1371/journal.pntd.0001637.

36. E. F. Briefer et al., "Mother goats do not forget their kids' calls," *Proceedings of the Royal Society: Biological Sciences,* 279 (2012), pp. 3749–3755. doi:10.1098/rspb.2012.0986.

37. R. E. Bonoan et al., "Vasculature of the hive: Heat dissipation in the honey bee (*Apis mellifera*) hive," *Die Naturwissenschaften,* 101 (2014), pp. 459–465. doi:10.1007/s00114-014-1174-2.

38. N. D. Volkow et al., “Effects of cell phone radiofrequency signal exposure on brain glucose metabolism,” *Journal of the American Medical Association,* 305 (2011), pp. 808–814.

39. NBC news report aired April 25, 2007. The study had been published in E. Lipov, S. Lipov, and J. T. Stark, “Stellate ganglion blockade provides relief from menopausal hot flashes: A case report series,” *Journal of Women’s Health,* 14 (2005), pp. 737–741.

40. M.-F. Yueh et al., “The commonly used antimicrobial additive triclosan is a liver tumor promoter,” *Proceedings of the National Academy of Sciences*, 111 (2014), pp. 17200–17205. doi:10.1073/pnas.1419119111.

41. R. F. Redberg, “Vitamin E and cardiovascular health,” *Journal of the American Medical Association,* 294 (2005), pp. 107–109.

42. R. C. Shelton et al., “Effectiveness of St. John’s wort in major depression,” *Journal of the American Medical Association*, 285 (2001), pp. 1978–1986.

43. A. Berende et al., “Randomized trial of longer-term therapy for symptoms attributed to Lyme disease,” *New England Journal of Medicine,* 374 (2016), pp. 1209–1220. doi:10.1056/NEJMoa1505425.

44. J. N. Kheir et al., “Oxygen gas-filled microparticles provide intravenous oxygen delivery,” *Science Translational Medicine*, 4 (2012), pp. 140ra88. doi:10.1126/scitranslmed.3003679.

45. “The notice of proposed rulemaking (NPRM) for revisions to the Common Rule: Summary of proposed changes relevant to genomics research,” NIH, 2015, www.genome.gov.

Chapter 8

1. L. H. Yang, “Periodical cicadas as resource pulses in North American forests,” *Science,* 306 (2004), pp. 1565–1567. The data are simulated Normal values that match the means and standard deviations reported in this article.

2. See Note 2 for Chapter 1.

3. “Behavioral health barometer: United States, 2015,” Substance Abuse and Mental Health Services Administration, www.samhsa.gov.

4. From the Nenana Ice Classic website, www.nenanaakiceclassic.com.

See R. Sagarin and F. Micheli, “Climate change in nontraditional data sets,” *Science*, 294 (2001), p. 811, for a careful discussion.

5. A. Montagne et al., “Blood–brain barrier breakdown in the aging human hippocampus,” *Neuron*, 85 (2015), pp. 296–302. doi:10.1016/j.neuron.2014.12.032.

6. A. S. Banks et al., “Juvenile hallux abducto valgus association with metatarsus adductus,” *Journal of the American Podiatric Medical Association*, 84 (1994), pp. 219–224.

7. See Note 16 for Chapter 5.

8. J. A. Martin et al., “Births: Final data for 2010,” *National Vital Statistics Reports*, 61 (2012).

9. R. Sihvonen et al., “Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear,” *New England Journal of Medicine*, 369 (2013), pp. 2515–2524. doi:10.1056/NEJMoa1305189.

10. Data from a plot in J. P. Rauschecker, B. Tian, and M. Hauser, “Processing of complex sounds in the macaque nonprimary auditory cortex,” *Science*, 268 (1995), pp. 111–114. The paper states that there are n=41$n = 41$ observations, but only n=37$n = 37$ can be read accurately from the plot.

11. B. Deer, “How the case against the MMR vaccine was fixed,” *BMJ*, 342 (2011), c5347. doi:10.1136/bmj.c5347.

12. See “Vaccine safety,” CDC, www.cdc.gov/vaccinesafety/concerns/autism.html.

13. See Note 10 for Chapter 7.

14. A. K. Davis and M. T. Holden, “Measuring intraspecific variation in flight-related morphology of monarch butterflies (*Danaus plexippus*): Which sex has the best flying gear?,”*Journal of Insects*, 2015 (2015), Article 591705. doi:10.1155/2015/591705.

15. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by A. Garth Fisher of Brigham Young University.

16. S. Alvarado et al., “Epigenetic variation in the *Egfr* gene generates quantitative variation in a complex trait in ants,” *Nature Communications*, 6 (2015), pp. 1–9. doi:10.1038/ncomms7513.

17. See Note 26 for Chapter 4.

Chapter 9

1. National Vital Statistics Reports, available at www.cdc.gov/nchs/products/nvsr.htm.
2. IPCC AR5 Working Group I, “Climate change 2013: The physical science basis,” www.ipcc.ch/pdf/unfccc/cop19/3_gregory13sbsta.pdf.
3. From the American Red Cross website, www.givelife2.org/aboutblood/bloodtypes.asp. There are some other blood types besides the eight main types, but they are extremely rare.
4. Florida Department of Health, Bureau of Epidemiology: “Animal rabies confirmed cases by county and animal type, January–December 2016,” www.floridahealth.gov.
5. G. M. Strain, “Deafness prevalence and pigmentation and gender associations in dog breeds at risk,” *Veterinary Journal*, 167 (2004), pp. 23–32.
6. Discrete sample spaces are most often, but not necessarily, a finite list of integers or individual outcomes. We will see in Chapter 12 the example of the Poisson distribution, a discrete distribution for X={0,1,2,...}.$X = \{0, 1, 2, \ldots\}$.
7. J. S. Schiller et al., “Summary health statistics for U.S. adults: National Health Interview Survey, 2011,” *Vital Health Statistics*, 10 (2012), Table 31.
8. 2004 *Statistical Abstract of the United States*, Table 193.
9. “Nearly half of Americans drink soda daily,” Gallup, July 23, 2012.
10. National Center for Health Statistics, “Health, United States, 2005,” p. 38.
11. *Kids and Media*, Kaiser Family Foundation Report, 1999.
12. The computations for the variance can often be faster by using an equivalent formula (where EE stands for “expected value,” or mean):

σ2=∑xi2*P(X=xi)-μ2=E(X2)-E(X)2

$$\sigma^2 = \sum x_i^2 * P(X = x_i) - \mu^2 = E(X^2) - E(X)^2$$

13. Data from the National Youth Risk Behavior Survey can be found on

the CDC website, www.cdc.gov/HealthyYouth/yrbs/data.

14. A. Leizorovicz et al., “Randomized, placebo-controlled trial of dalteparin for the prevention of venous thromboembolism in acutely ill medical patients,” *Circulation*, 110 (2004), pp. 874–879.

15. E. F. Dunne et al., “Prevalence of HPV infection among females in the United States,” *Journal of the American Medical Association*, 297 (2007), pp. 813–819.

16. G. Dezecache and R. I. M. Dunbar, “Sharing the joke: The size of natural laughter groups,” *Evolution and Human Behavior*, 33 (2012), pp. 775–779. doi:10.1016/j.evolhumbehav.2012.07.002.

17. J. Bose et al., “Key substance use and mental health indicators in the United States: Results from the 2015 National Survey on Drug Use and Health,” (2016), www.samhsa.gov.

18. See Note 25 for Chapter 1.

19. “Racial and ethnic distribution of ABO blood types,” www.bloodbook.com.

20. “Phenylketonuria,” *Free Health Encyclopedia*, faqs.org/health.

21. “Beer reigns as Americans’ preferred alcoholic beverage,” Gallup, August 3, 2016.

22. Health Resources and Services Administration, “Providing HIV/AIDS care in a changing environment: Hepatitis C and HIV co-infection,” September 2003.

23. “National diabetes statistics report, 2014,” CDC, www.cdc.gov/diabetes.

24. E. S. Siris et al., “Identification and fracture outcomes of undiagnosed low bone mineral density in postmenopausal women: Results from the National Osteoporosis Risk Assessment,” *Journal of the American Medical Association*, 286 (2001), pp. 2815–2822.

25. A. N. Dey and B. Bloom, “Summary health statistics for U.S. children: National Health Interview Survey, 2003,” *Vital Health Statistics*, 10 (2005), Table 17.

26. “Reported STDs in the United States, 2014 national data for chlamydia, gonorrhea, and syphilis,” CDC Fact Sheet, www.cdc.gov/std.

27. R. N. Rosenfiel et al., “Comparative relationships among eye color, age, and sex in three North American populations of Cooper’s hawks,” *Wilson Bulletin*, 115 (2003), pp. 225–230.
28. “Revised estimates of mental illness from the National Survey on Drug Use and Health,” *NSDUH Report*, November 19, 2013.
29. National Cancer Institute, “Fact sheet: Probability of breast cancer in American women,” www.cancer.gov.
30. N. Eriksson et al., “A genetic variant near olfactory receptor genes influences cilantro preference,” arXiv:1209.2096v1 [q-bio.GN] (2012).

Chapter 10

1. See Note 3 for Chapter 9.
2. This is one of several tests discussed in Bernard M. Branson, “Rapid HIV testing: 2005 update,” a presentation by the CDC, at www.cdc.gov. The Malawi clinic result is reported by Bernard M. Branson, “Point-of-care rapid tests for HIV antibody,” *Journal of Laboratory Medicine*, 27 (2003), pp. 288–295.
3. From the Florida Department of Health website, www.doh.state.fl.us, Brain and Spinal Cord Injury Program.
4. T. A. Lieu et al., “Geographic clusters in underimmunization and vaccine refusal,” *Pediatrics*, 135 (2015), peds.2014-2715. doi:10.1542/peds.2014-2715.
5. See Note 17 for Chapter 4.
6. Chronic Kidney Disease Surveillance System, www.cdc.gov/ckd.
7. S. L. Hedden et al., “Behavioral health trends in the United States: Results from the 2014 National Survey on Drug Use and Health,” (2015), www.samhsa.gov.
8. See Note 5 for Chapter 9.
9. A. Jamal et al., “Current cigarette smoking among adults – United States, 2005–2014,” *Morbidity and Mortality Weekly Report*, 64 (2015), pp. 1233–1240.
10. Royal Statistical Society news release, “Royal Statistical Society concerned by issues raised in Sally Clark case,” October 23, 2001, www.rss.org.uk. For background, see an editorial and article in *The*

Economist, January 22, 2004, “The probability of injustice.”

11. See Note 23 for Chapter 9.

12. A. C. Allison and D. F. Clyde, “Malaria in African children with deficient erythrocyte dehydrogenase,” *British Medical Journal*, 1 (1961), pp. 1346–1349.

13. N. Whitehead, “When should surgeons stop operating?”*NPR*, June 18, 2015.

14. See Note 7.

15. L. Naldi et al., “Pigmentary traits, modalities of sun reaction, history of sunburns, and melanocytic nevi as risk factors for cutaneous malignant melanoma in the Italian population: Results of a collaborative case-control study,” *Cancer*, 88 (2000), pp. 2703–2710.

16. W. Uter et al., “Inter-relation between variables determining constitutional UV sensitivity in Caucasian children,” *Photodermatology, Photoimmunology and Photomedicine*, 20 (2004), pp. 9–13.

17. W. D. Mosher et al., “Intended and unintended births in the United States: 1982–2010.” *National Health Statistics Reports*, 55 (2012).

18. G. Gigerenzer and A. Edwards, “Simple tools for understanding risks: From innumeracy to insight,” *British Medical Journal*, 327 (2003), pp. 741–744.

19. See Note 29 for Chapter 9.

20. Different countries use different cutoff points for deciding whether to call a result positive or negative, and some countries have each mammogram reviewed by two different experts. Thus, sensitivity and specificity vary from country to country. See U.S. Preventive Services Task Force, Agency for Healthcare Research and Quality, “Screening for breast cancer: Recommendations and rationale,” February 2002, www.ahrq.gov.

21. A. B. Miller et al., “Twenty five year follow-up for breast cancer incidence and mortality of the Canadian National Breast Screening Study: randomised screening trial,” *BMJ*, 348 (2014), g366. doi:10.1136/bmj.g366; H. G. Welch and H. J. Passow, “Quantifying the benefits and harms of screening mammography,” *JAMA Internal*

Medicine, 174 (2014), pp. 448–454. doi:10.1001/jamainternmed.2013.13635.

22. J. Boncy et al., “Performance and utility of a rapid diagnostic test for cholera: Notes from Haiti,” *Diagnostic Microbiology and Infectious Disease*, 76 (2013), pp. 521–523. doi:10.1016/j.diagmicrobio.2013.03.010.

23. See Note 18.

24. A. A. Wright and I. T. Katz, “Home testing for HIV,” *New England Journal of Medicine*, 354 (2006), pp. 437–440. doi:10.1056/NEJMp058302; “FDA approves first over-the-counter home use HIV test kit,” FDA News Release, July 3, 2012.

25. “Breast guidelines test American tolerance for risk,” *Reuters*, November 20, 2009, www.reuters.com.

26. “Silicosis: From public menace to litigation target,” *NPR*, March 6, 2006, www.npr.org. The website also links to the judge’s filed opinion, “Silica products liability litigation,” MDL Docket No. 1553.

27. T. Laksanasopin et al., “A smartphone dongle for diagnosis of infectious diseases at the point of care,” *Science Translational Medicine*, 7 (2015), pp. 273re1. doi:10.1126/scitranslmed.aaa0056.

28. L. Newman et al., “Global estimates of syphilis in pregnancy and associated adverse outcomes: Analysis of multinational antenatal surveillance data,” *PLOS Medicine*, 10 (2013), e.1001396. doi:10.1371/journal.pmed.1001396.

29. See Note 30 for Chapter 9.

30. www.cdc.gov/cancer/lung/basic_info/risk_factors.htm.

31. Polydactyly is not reported systematically, so estimated rates vary. National Center for Health Statistics, Series 20, No. 31 (1996); “Supernumerary digit,” “Polydactyly of the foot,” *Medscape*, emedicine.medscape.com.

32. S. W. Guo and D. R. Reed, “The genetics of phenylthiocarbamide perception,” *Annals of Human Biology*, 28 (2001), pp. 111–142.

33. See Note 16.

34. M. McCulloch et al., “Diagnostic accuracy of canine scent detection in early- and late-stage lung and breast cancers,” *Integrative Cancer*

Therapies, 5 (2006), pp. 30–39.

35. Data provided by Patricia Heithaus and the Department of Biology at Kenyon College.

36. See Note 21 for Chapter 1.

37. See Note 24 for Chapter 9.

38. R. S. Gupta et al., “The prevalence, severity, and distribution of childhood food allergy in the United States,” *Pediatrics*, 128 (2011), peds.2011-0204. doi:10.1542/peds.2011-0204.

39. See Note 22 for Chapter 9.

40. National Tay-Sachs and Allied Diseases Association, www.ntsad.org.

41. S. D. Grosse et al., “Newborn screening for cystic fibrosis,” *Morbidity and Mortality Weekly Report*, October 15, 2004.

42. National Institutes of Health, National Digestive Diseases Information Clearinghouse, wrongdiagnosis.com.

43. B. S. Pyenson et al., “Offering lung cancer screening to high-risk Medicare beneficiaries saves lives and is cost effective: An actuarial analysis,” *American Health & Drug Benefits*, 7 (2014), pp. 272–282.

44. OraQuick In-Home HIV Test customer package insert information, downloaded from www.orasure.com.

45. *Should you get a PSA test?*, 2006 Kaiser Permanente pamphlet based on “Should you get a PSA test? A patient–doctor decision,” American Academy of Family Physicians, www.aafp.org.

Chapter 11

1. The equation for the density curve of a Normal distribution with mean μμ and standard deviation σσ is 1σ2πe-12(x-μσ)2.$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

2. M. Jakob, *Normal Values pocket*, Borm Bruckmeier Publishing, 2002.

3. From a figure in C. M. Bishop et al., “The roller coaster flight strategy of bar-headed geese conserves energy during Himalayan migrations,” *Science*, 347 (2015), pp. 250–254. doi:10.1126/science.1258732.

4. “Osteoporosis: Review of the evidence for prevention, diagnosis and treatment and cost-effective analysis,” *Osteoporosis International*, 8 (1998), Suppl. 4.

5. E. F. Eriksen, “Treatment of osteopenia,” *Reviews in Endocrine & Metabolic Disorders*, 13 (2012), pp. 209–223. doi:10.1007/s11154-011-9187-z.
6. J. Lyall et al., “Suppression of avian influenza transmission in genetically modified chickens,” *Science*, 331 (2011), pp. 223–226. doi:10.1126/science.1198020.
7. R. C. Nelson, C. M. Brooks, and N. L. Pike, “Biomechanical comparison of male and female distance runners,” in P. Milvy (ed.), *The Marathon: Physiological, Medical, Epidemiological, and Psychological Studies*, New York Academy of Sciences, 1977, pp. 793–807.
8. See Note 2.
9. See Note 2.
10. See Note 32 for Chapter 1.
11. See Note 2.
12. A. John et al., “Trafficlike collective movement of ants on trails: Absence of a jammed phase,” *Physical Review Letters*, 102 (2009), pp. 108001(4). doi:10.1103/PhysRevLett.102.108001.
13. C.K. Folland et al., “Observed climate variability and change,” (2001). In: “Climate change 2001: The scientific basis. Contribution of Working Group I to the third assessment report of the Intergovernmental Panel on Climate Change,” Cambridge University Press. The graph appears on page 155.
14. Approximated values from Figure 6 in J. Hansen et al., “Global temperature in 2015,” (2016), www.columbia.edu/~jeh1/mailings/2016/20160120_Temperature2015.pdf.
15. See Note 10 for Chapter 2.
16. See Note 8 for Chapter 3.

Chapter 12

1. See Note 3 for Chapter 9.
2. B. Bekan Homawoo et al., “Chlamydia mapping by New York State school district boundaries, excluding New York City (2006–2007),” New York Bureau of STD Prevention and Epidemiology,

www.health.ny.gov.

3. S. Kidd et al., “Cephalosporin resistant *Neisseria gonorrhoeae* public health response plan,” CDC (2012).
4. R. D. Kirkcaldy et al., “*Neisseria gonorrhoeae* antimicrobial susceptibility surveillance – The Gonococcal Isolate Surveillance Project, 27 sites, United States, 2014,” *MMWR Surveillance Summaries*, 65 (2016), pp. 1–19. doi:10.15585/mmwr.ss6507a1.
5. Nalini Ranjit et al., “Contraceptive failure in the first two years of use: Differences across socioeconomic subgroups,” *Family Planning Perspectives,* 33 (2001), pp. 19–27.
6. M. Anderson and K. Cardoza, “Mental health in schools: A hidden crisis affecting millions of students,” *NPR,* August 31, 2016.
7. *Morbidity and Mortality Weekly Report,* March 30, 2006, www.cdc.gov.
8. See Note 7.
9. C. T. Mai et al., “Selected birth defects data from population-based birth defects surveillance programs in the United States, 2006 to 2010: Featuring trisomy conditions,” *Birth Defects Research*, 97 (2013), pp. 709–725.
10. CDC, Division of Bacterial and Mycotic Diseases, “Typhoid fever,” www.cdc.gov.
11. www.cdc.gov/nchs/fastats/emergency-department.htm.
12. See Note 4 for Chapter 10.
13. United Nations, “Human development report 2005,” Table 9: Leading global health crises and risks.
14. Associated Press, December 9, 2007, retrieved from www.msnbc.msn.com.
15. Florida Department of Health:“20-year animal rabies summary by species,” www.floridahealth.gov.
16. *Morbidity and Mortality Weekly Report*, October 14, 1994, www.cdc.gov.
17. *Morbidity and Mortality Weekly Report*, February 12, 2010, www.cdc.gov.
18. R. D. Andrews and A. Holland, “Oklahoma geological survey:

Statement on Oklahoma seismicity,” April 21, 2015.

19. Data available at earthquakes.ok.gov.

20. See Note 18.

21. earthquake.usgs.gov/learn/topics/megaqk_facts_fantasy.php.

22. See Note 24 for Chapter 9.

23. See Note 7 for Chapter 9.

Chapter 13

1. Graph of cigarette use among adults (Behavior Risk Factor Surveillance System), chronicdata.cdc.gov/health-area/tobacco-use.

2. B. E. Hamilton et al., “Births: Final data for 2014,” *National vital statistics reports*, 64 (2015), pp. 1–64.

3. “Percentage of pack-a-day smokers hits record low in U.S.,” Gallup, August 8, 2016.

4. Strictly speaking, the formula σ⁄n$\sigma/\sqrt{n}$ for the standard deviation of x-$\bar{x}$ assumes that we draw an SRS of size *n* from an *infinite* population. If the population has finite size N,N, this standard deviation is multiplied by 1-(n-1)⁄(N-1).$\sqrt{1-(n-1)/(N-1)}$. This “finite population correction” approaches 1 as NN increases. When the population is at least 20 times as large as the sample, the correction factor is between about 0.97 and 1. It is reasonable to use the simpler form σ⁄n$\sigma/\sqrt{n}$ in these settings.

5. This is true as long as the population has a finite standard deviation σ.
σ.

6. Strictly speaking, the formula p(1-p)⁄n$\sqrt{p(1-p)/n}$ for the standard deviation of p ^$\hat{p}$ assumes that we draw an SRS of size *n* from an *infinite* population. If the population has finite size N,N, this standard deviation is multiplied by 1-(n-1)⁄(N-1).$\sqrt{1-(n-1)/(N-1)}$. This “finite population correction” approaches 1 as NN increases. When the population is at least 20 times as large as the sample, the correction factor is between about 0.97 and 1. It is reasonable to use the simpler form p(1-p)⁄n$\sqrt{p(1-p)/n}$ in these settings.

7. See Note 4 for Chapter 12.

8. See Note 5 for Chapter 12.

Chapter 14

1. This value of σσ is based on the growth curves developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).
2. Data provided by Darlene Gordon, Purdue University.
3. See Note 21 for Chapter 2.
4. See Note 2 for Chapter 11.
5. M. Park et al., “Recycling endosomes supply AMPA receptors for LTP,” *Science*, 305 (2004), pp. 1972–1975.
6. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by A. L. Shoemaker of Calvin College. The original data were published in P. A. Mackowiak et al., “A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich,” *Journal of the American Medical Association*, 268 (1992), pp. 1578–1580.
7. J.-C. Lin et al., “Association between chronic hepatitis C virus infection and bone mineral density,” *Calcified Tissue International*, 91 (2012), pp. 423–429. doi:10.1007/s00223-012-9653-y.
8. We defined the null hypothesis as a specific statement (typically of “no effect”) about the exact value of the population parameter. However, you may sometimes see a null hypothesis expressed as the complement of the alternative hypothesis. The difference is noticeable only for one-sided tests. Consider Example 14.5 about inorganic phosphorus levels in an elderly population. Instead of H0:μ=1.2,$H_0 : \mu = 1.2,$ you would see H0:μ≥1.2,$H_0 : \mu \geq 1.2,$ because the alternative is Ha:μ<1.2.$H_a : \mu < 1.2.$ This approach certainly has a logic to it (of course, it is not impossible for the population mean inorganic phosphorus level to be greater than 1.2 mmol/l). However, this less specific null hypothesis is not the hypothesis used to compute the test *P*-value—the probability, computed supposing H0H_0 to be true, that the test statistic would take a value at least as extreme as that actually observed.
9. C. D. Gardner et al., “Effect of raw garlic vs. commercial garlic supplements on plasma lipid concentrations in adults with moderate hypercholesterolemia: A randomized clinical trial,” *Archives of*

Internal Medicine, 167 (2007), pp. 346–353.

10. W. H. Burt and R. P. Grossenheider, *A Field Guide to the Mammals: Field Marks of All North American Species Found North of Mexico*, 3rd ed., Houghton Mifflin, 1976.
11. Avandia (rosiglitazone maleate) prescribing information can be found at www.avandia.com.
12. Self-reported heights (in inches); course project.
13. Information retrieved from the U.S. Department of Agriculture website, www.ams.usda.gov/howtobuy/eggs.htm.
14. E. L. Kenney et al., “Prevalence of inadequate hydration among US children and disparities by gender and race/ethnicity: National Health and Nutrition Examination Survey, 2009–2012,” *American Journal of Public Health*, 105 (2015), pp. e113–e118. doi:10.2105/AJPH.2015.302572.
15. E. R. Spangenberg et al., “Effects of gender-congruent ambient scent on approach and avoidance behaviors in a retail store,” *Seventh Annual Retail Strategy and Consumer Decision Research Symposium*, 2004.

Chapter 15

1. You can find detailed information on the CDC website, www.cdc.gov, about the sampling design and analytic guidelines. Here is an excerpt: “Because NHANES is a complex probability sample, analytic approaches based on data from simple random samples are usually not appropriate. Ignoring the complex design can lead to biased estimates and overstated significance levels. Sample weights and the stratification and clustering of the design must be incorporated into an analysis to get proper estimates and standard errors of estimates.”
2. “Regulation of sugar-sweetened beverages,” *New England Journal of Medicine*. doi:10.1056/NEJMclde1210278.
3. See Note 2 for Chapter 1.
4. G. I. Pinkhasov et al., “Complications following prostate needle biopsy requiring hospital admission or emergency department visits – experience from 1000 consecutive cases,” *BJU International*, 110 (2012), pp. 369–374. doi:10.1111/j.1464-410X.2011.10926.x.
5. J. E. Guillory et al., “Editorial expression of concern: Experimental

evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, 111 (2014), pp. 10779–10779. doi:10.1073/pnas. 1412469111. R. Morin, "Facebook's experiment causes a lot of fuss for little result," Pew Research Center, July 2, 2014.

6. R. A. Laven and C. T. Livesey, "The effect of housing and methionine intake on hoof horn hemorrhages in primiparous lactating Holstein cows," *Journal of Dairy Science*, 87 (2004), pp. 1015–1023.

7. P. Weiner et al. "The cumulative effect of long-acting bronchodilators, exercise, and inspiratory muscle training on the perception of dyspnea in patients with advanced COPD," *Chest*, 118 (2000), pp. 672–678.

8. R. E. Kass et al., "Ten simple rules for effective statistical practice," *PLOS Computational Biology*, 12 (2016), e1004961. doi:10.1371/journal.pcbi.1004961.

9. B. J. Marshall and J. R. Warren, "Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration," *Lancet*, 1 (1984), pp. 1311–1315; B. J. Marshall et al., "A prospective double-blind trial of duodenal ulcer relapse after eradication of *Campylobacter pylori*," *Lancet*, 2 (1988), pp. 1437–1442; B. J. Marshall et al., "Urea protects *Helicobacter* (*Campylobacter*) *pylori* from the bactericidal effect of acid," *Gastroenterology*, 99 (1990), pp. 697–702.

10. E. M. Barsamian, "The rise and fall of internal mammary artery ligation," in J. P. Bunker, B. A. Barnes, and F. Mosteller (eds.), *Costs, Risks, and Benefits of Surgery*, Oxford University Press, 1977, pp. 212–220.

11. See Note 17 for Chapter 7.

12. G. Gregoratos et al., "ACC/AHA guidelines for implantation of cardiac pacemakers and antiarrhythmia devices: Executive summary," *Circulation*, 97 (1998), pp. 1325–1335.

13. K. Foden-Vencil, "Portland, Ore., becomes latest fluoride battleground," NPR, September 12, 2012.

14. From Table 3 of Belsomra's "Highlights of prescribing information," retrieved from Merck's website, www.merck.com.

15. INTERPHONE Study Group, "Brain tumour risk in relation to mobile telephone use: Results of the INTERPHONE international case-control

study," *International Journal of Epidemiology,* 39 (2010), pp. 675–694. doi:10.1093/ije/dyq079.

16. See Note 10 for Chapter 14.
17. C. Kopp et al., "Modulation of rhythmic brain activity by diazepam: GABAa receptor subtype and state specificity," *Proceedings of the National Academy of Sciences,* 101 (2004), pp. 3674–3679.
18. J. B. Moseley et al., "A controlled trial of arthroscopic surgery for osteoarthritis of the knee," *New England Journal of Medicine,* 347 (2002), pp. 81–88.

Chapter 16

1. See Note 2 for Chapter 13.
2. M. Hack et al., "Outcomes in young adulthood for very-low-birth-weight infants," *New England Journal of Medicine,* 346 (2002), pp. 149–157.
3. R. L. Wasserstein and N. A. Lazar, "The ASA's statement on *p*-values: Context, process, and purpose," *American Statistician,* 70 (2016), pp. 129–133. doi:10.1080/00031305.2016.1154108.
4. See Note 16 for Chapter 10.
5. Statistics Canada, Census at School, 2010–2011, www.censusatschool.ca/data-results/2010-2011.
6. Population base from the 2005 General Household Survey, www.statistics.gov.uk. Smoking data from Action on Smoking and Health, *Smoking and Health Inequality*, www.ash.org.uk.
7. www.drugabuse.gov.
8. F. J. Sulloway, "Birth order and intelligence," *Science,* 316 (2007), pp. 1711–1712. The Norwegian study is Petter Kristensen and Tor Bjerkedal, "Explaining the relation between birth order and intelligence," *Science,* 316 (2007), p. 1717.
9. J. F. Swain et al., "Comparison of the effects of oat bran and low-fiber wheat on serum lipoprotein levels and blood pressure," *New England Journal of Medicine,* 322 (1990), pp. 147–152.

Chapter 17

1. Note 1 for Chapter 13 explains the reason for this condition in the case of inference about a population mean. In practice, inference computations require that the population be much larger than the sample—say, at least 20 times as large. All of our examples and exercises satisfy this condition. Practical settings in which the sample is a large part of the population are rather rare and special, and we will not discuss them.
2. T. Hesselberg and F. Vollrath, “The mechanical properties of the non-sticky spiral in *Nephila* orb webs (Araneae, Nephilidae),” *Journal of Experimental Biology,* 215 (2012), pp. 3362–3369. doi:10.1242/jeb.068890.
3. The degrees of freedom for the one-sample t statistic come from the sample standard deviation s in the denominator of t. We saw in Chapter 2 (page 49) that s has n-1$n - 1$ degrees of freedom.
4. M. Mogri et al., “Testosterone concentrations in young pubertal and post-pubertal obese males,” *Clinical Endocrinology,* 78 (2013), pp. 593–599. doi:10.1111/cen. 12018.
5. K. L. Crossley et al., “What is an attractive body? Using an interactive 3D program to create the ideal body for you and your partner,” *PLOS ONE,* 7 (2012), e50601. doi:10.1371/journal.pone.0050601.
6. See Note 13 for Chapter 8.
7. Z. W. Liller, “Spatial variation in brook trout (*Salvelinus fontinalis*) population dynamics and juvenile recruitment potential in an Appalachian watershed,” MS thesis, West Virginia University, 2006.
8. L. B. Gordon et al., “Clinical trial of a farnesyltransferase inhibitor in children with Hutchinson-Gilford progeria syndrome,” *Proceedings of the National Academy of Sciences,* 109 (2012), pp. 16666–16671. doi:10.1073/pnas.1202529109.
9. M. A. Case et al., “Accuracy of smartphone applications and wearable devices for tracking physical activity data,” *Journal of American Medical Association,* 313 (2015), pp. 625–625. doi:10.1001/jama.2014.17841.
10. What matters is reducing the original data to a one-sample problem. Taking the difference between the two conditions is the most common approach. One could also, for instance, calculate the percent change for

each pair of data points.

11. E. A. Dennis et al., “Water consumption increases weight loss during a hypocaloric diet intervention in middle-aged and older adults,” *Obesity,* 18 (2010), pp. 300–307. doi:10.1038/oby.2009.235.

12. S. J. Russell et al., “Outpatient glycemic control with a bionic pancreas in type 1 diabetes,” *New England Journal of Medicine,* 371 (2014), pp. 313–325. doi:10.1056/NEJMoa1314474.

13. For a qualitative discussion explaining why skewness is the most serious violation of the Normal shape condition, see D. D. Boos and J. M. Hughes-Oliver, “How large does n have to be for the Z and t intervals?,” *American Statistician,* 54 (2000), pp. 121–128. Our recommendations are based on extensive computer work. See, for example, H. O. Posten, “The robustness of the one-sample t-test over the Pearson system,” *Journal of Statistical Computation and Simulation,* 9 (1979), pp. 133–149; and E. S. Pearson and N. W. Please, “Relation between the shape of population distribution and the robustness of four simple test statistics,” *Biometrika,* 62 (1975), pp. 223–241.

14. See Note 10 for Chapter 2.

15. See Note 1 for Chapter 2.

16. See Note 10 for Chapter 6.

17. M. Rundlöf et al., “Seed coating with a neonicotinoid insecticide negatively affects wild bees,” *Nature,* 521 (2015), pp. 77–80. doi:10.1038/nature14420.

18. R. Sakakibara et al., “Influence of body position on defecation in humans,” *Lower Urinary Tract Symptoms,* 2 (2010), pp. 16–21. doi:10.1111/j.1757-5672.2009.00057.x.

19. M. Anders, *ACE FitnessMatters,* July–August 2008, pp. 7–9.

20. See Note 14 for Chapter 7.

21. See Note 8 for Chapter 1.

22. Data from the “wine” database in the archive of machine-learning databases at the University of California, Irvine, ftp.ics.uci.edu/pub/machine-learning-databases.

23. We thank Rudi Berkelhamer of the University of California at Irvine for the data. These data are part of a larger set collected for an undergraduate lab exercise in scientific methods.

24. See Note 2.

25. See Note 6 for Chapter 2.

26. T. W. McDowell, “An evaluation of vibration and other effects on the accuracy of grip and push force recall,” PhD thesis, West Virginia University, 2006.

27. Data for Cohort 2 in R. A. Morgan et al., “Cancer regression in patients after transfer of genetically engineered lymphocytes,” *Science,* 314 (2006), pp. 126–129. The doubling-time data are given in the paper, and the immune response data appear in the supplementary online material.

28. From Table S2 in the online supplement to A. Dell’Anno and R. Danovaro, “Extracellular DNA plays a key role in deep-sea ecosystem functioning,” *Science,* 309 (2005), p. 2179.

29. E. G. van den Heuvel et al., “Oligofructose stimulates calcium absorption in adolescents,” *American Journal of Clinical Nutrition,* 69 (1999), pp. 544–548.

30. R. de la Fuente-Fernandez et al., “Expectation and dopamine release: Mechanism of the placebo effect in Parkinson’s disease,” *Science,* 293 (2001), pp. 1164–1166.

31. See Note 14 for Chapter 2.

32. A. P. Melis, B. Hare, and M. Tomasello, “Chimpanzees recruit the best collaborators,” *Science,* 311 (2006), pp. 1297–1300. A Normal quantile plot does not show major lack of Normality, and a saddlepoint approximation that allows for skew gives P=0.0039.$P = 0.0039.$ Thus, the *t* test is reasonably accurate despite the skew and small sample size.

33. L. Yuh, “A biopharmaceutical example for undergraduate students,” manuscript.

34. Data courtesy of Brad Hughes, Department of Ecology and Evolutionary Biology, University of California at Irvine.

Chapter 18

1. See Note 40 for Chapter 7.

2. See Note 25 for Chapter 6.
3. See Note 18 for Chapter 2.
4. P. A. Miller et al., “Reducing the size of the human physiological blind spot through training,” *Current Biology*, 25 (2015), pp. R747–R748. doi:10.1016/j.cub.2015.07.026.
5. L. L. Imbach et al., “Sleep–wake disorders persist 18 months after traumatic brain injury but remain underrecognized,” *Neurology*, 86 (2016), pp. 1945–1950. doi:10.1212/WNL.0000000000002697.
6. See Note 6 for Chapter 11.
7. See Note 26 for Chapter 1.
8. See Note 8 for Chapter 5.
9. L. Schlaffke et al., “The brain’s dress code: How the dress allows to decode the neuronal pathway of an optical illusion,” *Cortex*, 73 (2015), pp. 271–275. doi:10.1016/j.cortex.2015.08.017.
10. See the extensive simulation studies in H. O. Posten, “The robustness of the two-sample *t*-test over the Pearson system,” *Journal of Statistical Computation and Simulation*, 6 (1978), pp. 295–311; and in H. O. Posten, H. Yeh, and D. B. Owen, “Robustness of the two-sample *t*-test under violations of the homogeneity assumption,” *Communications in Statistics*, 11 (1982), pp. 109–126.
11. J. Govero et al., “Zika virus infection damages the testes in mice,” *Nature*, 540 (2016), pp. 438–442. doi:10.1038/nature20556.
12. V. Ieto et al., “Effects oforopharyngeal exercises on snoring: A randomized trial,” *Chest*, 148 (2015), pp. 683–691. doi:10.1378/chest.14-2953.
13. See, for example, G. D. Ruxton, “The unequal variance *t*-test is an underused alternative to Student’s *t*-test and the Mann-Whitney *U* test,” *Behavioral Ecology*, 17 (2006), pp. 688–690.
14. The problem of comparing spreads is difficult even with advanced methods. Common distribution-free procedures do not offer a satisfactory alternative to the *F* test, because they are sensitive to unequal shapes when comparing two distributions. A recent survey of possible approaches is D. D. Boos and C. Brownie, “Comparing variances and other measures of dispersion,” *Statistical Science*, 19

(2005), pp. 571–578. See also Lewis H. Shoemaker, “Fixing the *F* test for equal variances,” *American Statistician*, 57 (2003), pp. 105–114, for adjustments to *F* that improve its robustness. The adjustments involve data-dependent degrees of freedom, similar in spirit to the two-sample *t* procedures described in this Chapter.

15. H. R. Kranzler et al., “Topiramate treatment for heavy drinkers: Moderation by a GRIK1 polymorphism,” *American Journal of Psychiatry*, 171 (2014), pp. 445–452. doi:10.1176/appi.ajp.2013.13081014.

16. J. M. Engelmann and E. Herrmann, “Chimpanzees trust their fFriends,” *Current Biology*, 26 (2016), pp. 252–256. doi:10.1016/j.cub.2015.11.037.

17. V. L. Haidinger, “Feather growth, molt and sexual selection in the longtailed finch, *Poephila acuticauda*,” MS thesis, University of California at Irvine, 1996.

18. E. R. Wald et al., “Addressing child nutrition and physical activity in the primary care setting,” Agency for Healthcare Research and Quality, 2010 conference, www.ahrq.gov.

19. Data provided by Darlene Gordon, Purdue University.

20. L. Z. Garamszegi et al., “Immune challenge mediates vocal communication in a passerine bird: an experiment,” *Behavioral Ecology*, 15 (2004), pp. 148–157.

21. Data provided by Corinne Lim, Purdue University, from a student project supervised by Professor Joseph Vanable.

22. A. Sekar et al., “Schizophrenia risk from complex variation of complement component 4,” *Nature*, 530 (2016), pp. 177–183. doi:10.1038/nature16549.

23. From the Avandia website, www.avandia.com, prescribing information.

24. D. Ziebolz et al., “Long-term effects of tongue piercing: A case control study,” *Clinical Oral Investigations*, 16 (2012), pp. 231–237. doi:10.1007/s00784-011-0510-6.

25. R. E. Oldham-Cooper et al., “Playing a computer game during lunch affects fullness, memory for lunch, and later snack intake,” *American*

Journal of Clinical Nutrition, 93 (2011), pp. 308–313.

26. Ola Andersson et al., “Effect of delayed cord clamping on neurodevelopment at 4 years of age: A randomized clinical trial,” *JAMA Pediatrics*, 169 (2015), pp. 631–638. doi:10.1001/jamapediatrics.2015.0358.

27. See Note 5 for Chapter 7.

28. Data provided by Marigene Arnold, Kalamazoo College.

29. *Black Newsprint Inks*, National Toicology Program Toxicity Report Series No. 17, ntp.niehs.nih.gov.

30. C. Vallbona et al., “Response of pain to static magnetic fields in postpolio patients: A double-blind pilot study,” *Archives of Physical Medicine and Rehabilitation*, 78 (1997), pp. 1200–1203.

31. See Note 8 for Chapter 3.

Chapter 19

1. See Note 15 for Chapter 9.

2. N. Lindson-Hawley et al., “Gradual versus abrupt smoking cessation: A randomized, controlled noninferiority trial,” *Annals of Internal Medicine*, 164 (2016), pp. 585–592. doi:10.7326/M14-2805.

3. See Note 4 for Chapter 13.

4. Retrieved from the Florida Department of Health’s website, www.doh.state.fl.us, Epidemiology, Antibiotic resistance.

5. R. Stein, “Scientists see upside and downside of sequencing their own genes, *NPR*, September 19, 2012.

6. This rule of thumb is based on a study of computational results in the papers cited in Note 8 and in discussion with Alan Agresti. We strongly recommend using the plus four interval. See Note 9.

7. To address major issues of undercoverage and low response rates, pollsters may weight the answers of different groups to make the sample more representative of the target population. This approach affects the computation of the margin of error, however, resulting in a wider confidence interval. Gallup, for instance, routinely states that “All reported margins of sampling error include computed design effects for weighting.” Weighting and its corresponding effect on the margin of error are beyond the scope of an introductory statistics textbook.

8. J. S. M. Peiris et al., “Prospective study of the clinical progression and viral load of SARS associated coronavirus pneumonia in a community outbreak,” available on the website of the World Health Organization, www.who.int.

9. This interval is proposed by A. Agresti and B. A. Coull, “Approximate is better than ‘exact’ for interval estimation of binomial proportions,” *American Statistician,* 52 (1998), pp. 119–126. Note in particular that the plus four interval is often more accurate than the Clopper-Pearson “exact interval” based on the binomial distribution of the sample count and implemented by, for example, Minitab.

There are several even more accurate but considerably more complex intervals for *p* that might be used in professional practice. See L. D. Brown, T. Cai, and A. DasGupta, “Interval estimation for a binomial proportion,” *Statistical Science,* 16 (2001), pp. 101–133. A detailed theoretical study that uncovers the reason why the large-sample interval is inaccurate is L. D. Brown, T. Cai, and A. DasGupta, “Confidence intervals for a binomial proportion and asymptotic expansions,” *Annals of Statistics,* 30 (2002), pp. 160–201.

10. D. Fergusson et al., “Turning a blind eye: The success of blinding reported in a random sample of randomised, placebo-controlled trials,” *British Medical Journal,* 328 (2004), pp. 432–436.

11. From A. Agresti and B. Caffo, “Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures,” *American Statistician,* 45 (2000), pp. 280–288. When can the plus four interval be safely used? The answer depends on just how much accuracy you insist on. Brown and coauthors (see Note 9) recommend $n \geq 40$.$n \geq 40$. Agresti and Coull demonstrate that performance is almost always satisfactory in their eyes when $n \geq 5$.$n \geq 5$. Our rule of thumb of $n \geq 10$$n \geq 10$ allows for confidence levels *C* other than 95% and fits our philosophy of not insisting on more exact results than practice requires. The big point is that plus four is much more accurate than the standard interval for most values of *p* and all but very large *n*.

12. In fact, *P*-values for two-sided tests are more accurate than those for one-sided tests. Our rule of thumb is a compromise to avoid the confusion of too many rules.

13. G. Ribak et al., “Adaptive aerial righting during the escape dropping of wingless pea aphids,” *Current Biology*, 23 (2013), pp. R102–R103. doi:10.1016/j.cub.2012.12.010.

14. M. A. Doft et al, “The Newborn Butterfly Project,” *Plastic and Reconstructive Surgery*, 135 (2015), pp. 577e–583e. doi:10.1097/PRS.0000000000000999.

15. S. S. Magill et al, “Multistate point-prevalence survey of health care–associated infections,” *New England Journal of Medicine*, 370 (2014), pp. 1198–1208. doi:10.1056/NEJMoa1306801.

16. D. M. Robertson and H. D. Cavanagh, “Non-compliance with contact lens wear and care practices: A comparative analysis,” *Optometry and Vision Science*, 88 (2011), pp. 1402–1408. doi:10.1097/OPX.0b013e3182333cf9.

17. S. C. Markt et al, “Sniffing out significant ’pee values’: Genome wide association study of asparagus anosmia,” *BMJ*, 2016, i6071. doi:10.1136/bmj.i6071.

18. C. M. Willis et al., “Olfactory detection of human bladder cancer by dogs: Proof of principle study,” *British Medical Journal*, 329 (2004), pp. 712–717.

19. “Lead in DC drinking water,” U.S. Environmental Protection Agency, www.epa.gov/dclead.

20. A. Todorov et al., “Inferences of competence from faces predict election outcomes,” *Science*, 308 (2005), pp. 1623–1626.

21. See Note 30 for Chapter 9.

22. See Note 7 for Chapter 1.

23. L. G. M. Bode, “Preventing surgical-site unfections in nasal carriers of *Staphylococcus aureus*,” *New England Journal of Medicine*, 362 (2010), pp. 9–17.

24. See Note 7 for Chapter 9.

Chapter 20

1. See Note 2 for Chapter 19.
2. See Note 10 for Chapter 5.
3. The plus four method was developed by Alan Agresti and Brian Caffo. See Notes 9 and 11 for Chapter 19.
4. K. F. Bieniek et al, "Chronic traumatic encephalopathy pathology in a neurodegenerative disorders brain bank," *Acta Neuropathologica*, 130 (2015), pp. 877–889. doi:10.1007/s00401-015-1502-4.
5. NCCAM website (nccam.nih.gov/health/echinacea/) and R. B. Turner et al., "An evaluation of *Echinacea angustifolia* in experimental rhinovirus infections," *New England Journal of Medicine*, 353 (2005), pp. 341–348.
6. From the Gardasil website (www.gardasil.com), Health Care Professionals, Efficacy Data.
7. "HIV vaccine regimen demonstrates modest preventive effect in Thailand clinical study," *NIH News*, September 24, 2009, www.niaid.nih.gov.
8. The proper count requirements for the two-sample proportions test of hypothesis should theoretically be about expected counts, as for the one-sample test (Chapter 19) and the chi-square tests (Chapters 21 and 22). However, the expected counts requirements become somewhat cumbersome for the two-sample procedure, so we provide a simpler alternative focused on actual counts. The reason for this decision is to avoid drowning students in noncritical computational details. Here are the requirements based on expected counts:

   If we have two samples of sizes n1n_1 and n2n_2 such that N=n1+n2, $N = n_1 + n_2$, and if we call SS the total number of successes and *F* the total number of failures in both samples combined, then each of the following four expected counts should be greater than 5: n1S⁄N,$n_1 S/N$, n2S⁄N,$n_2 S/N$, n1F⁄N,$n_1 F/N$, and n2F⁄N.$n_2 F/N$. This implies that the total number of successes and the total number of failures should each be equal to or greater than [5(n1+n2)⁄(larger of n1 or n2)]. $[5\,(n_1 + n_2)/(\text{larger of } n_1 \text{ or } n_2)]$. The sample counts requirements described in the text tend to be more conservative than the requirements using expected counts, especially when only one of the two samples is very small.

9. K. Kawaza et al., “Efficacy of a low-cost bubble CPAP system in treatment of respiratory distress in a neonatal ward in Malawi,” *PLOS ONE,* 9 (2014), e86327. doi:10.1371/journal.pone.0086327.

10. Steering Committee of the Physicians’ Health Study Research Group, “Final report on the aspirin component of the ongoing Physicians’ Health Study,” *New England Journal of Medicine*, 321 (1989), pp. 129–135.

11. E. de Jonge et al., “Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: A randomised controlled trial,” *Lancet,* 362 (2003), pp. 1011–1016.

12. A. J. Wakefield et al., “Ileal lymphoid nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children,” *Lancet,* 351 (1998), pp. 637–641 [retracted].

13. See Note 11 for Chapter 8.

14. See Note 30 for Chapter 9.

15. J. E. Tcheng, “High-dose eptifibatide (Integrilin) in elective coronary stenting: Results of the ESPRIT trial; late-breaking trial,” presented at the American College of Cardiology 49th Annual Scientific Session, March 12–15, 2000, Anaheim, CA.

16. F. Gaudet et al., “Induction of tumors in mice by genomic hypomethylation,” *Science,* 300 (2003), pp. 489–492.

17. F. Lloret et al., “Fire and resprouting in Mediterranean ecosystems: Insights from an external biogeographical region, the Mexican shrubland,” *American Journal of Botany*, 88 (1999), pp. 1655–1661.

18. See Note 2 for Chapter 5.

19. See Note 9 for Chapter 8.

20. See Note 23 for Chapter 19.

21. K. deWeber et al., “Knuckle cracking and hand osteoarthritis,” *Journal of the American Board of Family Medicine,* 24 (2011), pp. 169–174. doi:10.3122/jabfm.2011.02.100156.

22. See Note 3 for Chapter 5.

23. R. L. Prentice et al., “Low-fat dietary pattern and risk of invasive breast cancer,” *Journal of the American Medical Association,* 295 (2006), pp. 629–642.

24. E. Azziz-Baumgartner et al., “Case-control study of an acute aflatoxicosis outbreak, Kenya, 2004,” *Environmental Health Perspectives*, 113 (2005), pp. 1779–1783.

25. B. Guyuron et al., “A placebo-controlled surgical trial of the treatment of migraine headaches,” *Plastic and Reconstructive Surgery*, 124 (2009), pp. 461–468.

26. www.pfizerpro.com, Chantix, Efficacy of Chantix.

27. D. R. Focht III, C. Spicer, and M. P. Fairchok, “The efficacy of duct tape vs. cryotherapy in the treatment of *Verruca vulgaris* (the common wart),” *Archives of Pediatrics and Adolescent Medicine*, 156 (2002), pp. 971–974.

28. R. F. Adair and L. R. Holmgren, “Do drug samples influence resident prescribing behavior? A randomized trial,” *American Journal of Medicine*, 118 (2005), pp. 881–884.

29. “Toxicology and carcinogenesis studies of 60-Hz magnetic fields in F344/N rats and B6C3F1 mice (whole-body exposure studies),” TR-488, ntp.niehs.nih.gov.

30. G. Bellipanni et al., “Effects of melatonin in perimenopausal and menopausal women,” *New York Academy of Sciences*, 1057 (2005), pp. 393–402.

31. “ACE inhibitors as a new frontier in cardiovascular prevention,” report from the Heart Outcomes Prevention Evaluation trial, Population Health Research Institute, www.ccc.mcmaster.ca.

32. G. D’Souza et al., “Case-control study of human papillomavirus and oropharyngeal cancer,” *New England Journal of Medicine*, 356 (2007), pp. 1944–1956. doi:10.1056/NEJMoa065497.

Chapter 21

1. P. A. Kulakow, H. Hauptli, and S. K. Jain, “Genetics of grain amaranths,” *Journal of Heredity*, 76 (1985), pp. 27–30.

2. See Note 13 for Chapter 19.

3. Based on a news item in *Science*, 305 (2004), p. 1560. The study, by Daniel Klem, appeared in the *Wilson Journal*.

4. H. Wei et al., “A dominant major locus in chromosome 9 of rice

(*Oryza sativa* L.) confers tolerance to 48° C$48^{\circ}\mathrm{C}$ high temperature at seedling stage," *Journal of Heredity*, 104 (2013), pp. 287–294. doi:10.1093/jhered/ess103.

5. Many computer studies have examined the accuracy of critical values for X2.χ^2. Our guideline goes back to Cochran (1954). Later work has shown that it is often conservative in the sense that if the expected cell counts are all similar and the degrees of freedom exceed 1, the chi-square approximation works well for an average expected count as small as 1 or 2. Our guideline protects against dissimilar expected counts.

6. P. R. Arelli, L. D. Young, and V. C. Concibido, "Inheritance of resistance in soybean PI 567516C to LY1 nematode population infecting cv. Hartwig," *Euphytica*, 165 (2009), pp. 1–4.

7. G. Noé et al., "Contraceptive efficacy of emergency contraception with levonorgestrel given before or after ovulation," *Contraception*, 81 (2010), pp. 414–420. doi:10.1016/j.contraception.2009.12.015.

8. Levy et al., "Arm coordination in octopus crawling involves unique motor control strategies," *Current Biology*, 25 (2015), pp. 1–6. doi:10.1016/j.cub.2015.02.064.

9. See Note 20 for Chapter 4.

10. R. D. Jackson et al., "Calcium plus vitamin D supplementation and the risk of fractures," *New England Journal of Medicine*, 354 (2006), pp. 669–683.

11. J. L. Hoogland, "Multiple mating by female prairie dogs," *Animal Behavior*, 55 (1998), pp. 351–359.

12. E. Miura et al., "The balance between protein synthesis and degradation in chloroplasts determines leaf variegation in *Arabidopsis* yellow variegated mutants," *Plant Cell*, 19 (2007), pp. 1313–1328.

13. See Note 6 for Chapter 11.

14. W. Landauer and L. C. Dunnthe, "Frizzle characters of fowls: Its expression and inheritance," *Journal of Heredity*, 21 (1930), pp. 291–305.

15. V. Buonaccorsi and A. Skibiel, "A striking demonstration of the Poisson distribution," *Teaching Statistics*, 27 (2005), pp. 8–10.

Chapter 22

1. See Note 42 for Chapter 6.

2. C. Bullen et al., “Electronic cigarettes for smoking cessation: A randomized controlled trial,” *Lancet*, 382 (2013), pp. 1629–1637. doi:10.1016/S0140-6736(13)61842-5.

3. See Note 13 for Chapter 19.

4. There are many computer studies of the accuracy of critical values for X2.χ^2. Our guideline goes back to Cochran (1954). Later work has shown that it is often conservative in the sense that if the expected cell counts are all similar and the degrees of freedom are greater than 1, the chi-square approximation works well for an average expected count as small as 1 or 2. Our guideline protects against dissimilar expected counts. It has the added advantage that it is safe in the 2×2$2 \times 2$ case, where the chi-square approximation is least good. So our condition is helpful for beginners—there is no single condition that is not conservative and that applies to 2×2$2 \times 2$ and larger tables with similar and dissimilar expected cell counts. There are exact procedures that (with software) should be used for tables that do not satisfy our condition. See A. Agresti, “A survey of exact inference for contingency tables,” *Statistical Science*, 7 (1992), pp. 131–177.

5. Pew Research Center, October 2014, “Online harassment.”

6. See Note 4 for Chapter 20.

7. Joseph P. Broderick et al., “Endovascular therapy after intravenous t-PA versus t-PA alone for stroke,” *New England Journal of Medicine*, 368 (2013), pp. 893–903. doi:10.1056/NEJMoa1214300.

8. See Note 9 for Chapter 20.

9. A. Sood et al., “S-Adenosyl-L-Methionine (SAMe) for smoking abstinence: A randomized clinical trial,” *Journal of Alternative and Complementary Medicine*, 18 (2012), pp. 854–859. doi:10.1089/acm.2011.0462.

10. See Note 6 for Chapter 9.

11. See Note 2 for Chapter 5.

12. T. Kontiokari et al., “Randomised trial of cranberry–lingonberry juice and Lactobacillus GG drink for the prevention of urinary tract infections in women," *BMJ*, 322 (2001), pp. 1–5.

13. B. C. Coleman, "Study: Heart attack risk cut 74% by stress management," Associated Press dispatch appearing in the Lafayette, Indiana, *Journal and Courier*, October 20, 1997.
14. Results presented by S. Vasan and D. Ho at the AIDS Vaccine 2009 conference in Paris, France, www.vaccineenterprise.org/conference_archive/2009.
15. C. Davison Ankney, "Sex-ratio varies with egg sequence in lesser snow geese," *Auk*, 99 (1982), pp. 662–666.
16. See Note 32 for Chapter 20.
17. See Note 23 for Chapter 18.
18. www.epiduo.com, "Full prescribing information."
19. See Note 17 for Chapter 4.
20. See Note 19 for Chapter 5.
21. See Note 8 for Chapter 5, and nccam.nih.gov/health/echinacea/.
22. See Note 15 for Chapter 10.
23. See Note 29 for Chapter 20.

Chapter 23

1. See Note 5 for Chapter 8.
2. K. Lebret et al., "Grazing resistance allows bloom formation and may explain invasion success of *Gonyostomum semen*," *Limnology and Oceanography*, 57(2012), pp. 727–734. doi:10.4319/lo.2012.57.3.0727.
3. See Note 14 for Chapter 4.
4. B. A. O'Brien et al., "The effect of print size on reading speed in dyslexia," *Journal of Research in Reading*, 28 (2005), pp. 332–349.
5. See Note 26 for Chapter 17.
6. C. J. Stevens et al., "Impact of nitrogen deposition on the species richness of grasslands," *Science*, 303 (2004), pp. 1876–1879.
7. A. P. Moller, "Rapid change in nest size of a bird related to change in a secondary sexual character," *Behavioal Ecology*, 17 (2006), pp. 108–116.

8. See Note 14 for Chapter 3.
9. See Note 15 for Chapter 3.
10. See Note 17 for Chapter 4.
11. See Note 5 for Chapter 2.
12. From a graph in B.-E. Saether, S. Engen, and E. Mattysen, "Demographic characteristics and population dynamical patterns of solitary birds," *Science*, 295 (2002), pp. 2070–2073.
13. See Note 4 for Chapter 17.
14. N. Bahrami et al., "Subconcussive head impact exposure and white matter tract changes over a single season of youth football," *Radiology*, 281 (2016), pp. 919–926. doi:10.1148/radiol.2016160564.
15. See Note 9 for Chapter 3.

Chapter 24

1. P. M. Johnson and P. J. Kenny, "Dopamine D2 receptors in addiction-like reward dysfunction and compulsive eating in obese rats," *Nature Neuroscience*, 13 (2010), pp. 635–641. doi:10.1038/nn.2519.
2. Based on the online supplement to P. J. Shaw et al., "Correlates of sleep and waking in *Drosophila melanogaster*," *Science*, 287 (2000), pp. 1834–1837.
3. K. A. Scott et al., "Cannabinoid activity in an orthotopic murine glioma model," *Molecular Cancer Therapeutics*, 13 (2014), pp. 2955–2967. doi:10.1158/1535-7163.MCT-14-0402.
4. See Note 19 for Chapter 2.
5. M. Okla et al., "Ellagic acid modulates lipid accumulation in primary human adipocytes and human hepatoma Huh7 cells via discrete mechanisms," *Journal of Nutritional Biochemistry*, 26 (2015), pp. 82–90. doi:10.1016/j.jnutbio.2014.09.010.
6. N. A. Sherif et al., "Detection of cotinine in neonate meconium as a marker for nicotine exposure in utero," *Eastern Mediterranean Health Journal*, 10 (2004), pp. 96–105.
7. C. A. Slentz et al., "The effects of aerobic versus resistance training on visceral and liver fat stores, liver enzymes and insulin resistance by HOMA in overweight adults from STRRIDE AT/RT: A randomized

trial," *American Journal of Physiology: Endocrinology and Metabolism,* 301 (2011), E1033–E1039. doi:10.1152/ajpendo.00291.2011.

8. C. D. Johnston et al., "Bonding to molars: The effect of etch time (an in-vitro study)," *European Journal of Orthodontics,* 20 (1998), pp. 195–199.

9. A. Adan and J. M. Serra-Grabulosa, "Effects of caffeine and glucose, alone and combined, on cognitive performance," *Human Psychopharmacology: Clinical and Experimental,* 25 (2010), pp. 310–317. doi:10.1002/hup.1115.

10. A. C. St-Pierre et al., "Insulin resistance syndrome, body mass index, and the risk of ischemic heart disease," *Canadian Medical Association Journal,* 172 (2005), pp. 1301–1305.

11. R. J. Safran et al., "Dynamic paternity allocation as a function of male plumage color in barn swallows," *Science,* 209 (2005), pp. 2210–2212.

12. Data from the online supplement to A. Kessler and I. T. Baldwin, "Defensive function of herbivore-induced plant volatile emissions in nature," *Science,* 291 (2001), pp. 2141–2144.

13. See Note 8 for Chapter 8.

14. M. Koch et al., "Hypothalamic POMC neurons promote cannabinoid-induced feeding," *Nature,* 519 (2015), pp. 45–50. doi:10.1038/nature14260.

15. See Note 34 for Chapter 17.

16. B. Stricanne, "Etudes d'integration multisensorielles dans la voie visuelle occipito-parietale du primate," PhD thesis, Université Paris VI, 1996.

17. J. T. Ngai and D. S. Srivastava, "Predators accelerate nutrient cycling in a bromeliad ecosystem," *Science,* 314 (2006), p. 963. We thank Jacqueline Ngai for providing the data.

18. P. S. Gupta, "Reaction of plants to the density of soil," *Journal of Ecology,* 21 (1933), pp. 452–474.

19. A. Alecu et al., "Effect of the homeopathic remedies *Arnica montana* and *Staphisagria* on the time of healing of surgical wounds," *Cultura Homeopática,* 20 (2007), pp. 19–21.

20. M. R. McClung et al., “Denosumab in postmenopausal women with low bone mineral density,” *New England Journal of Medicine*, 354 (2006), pp. 821–831.

21. We thank Rudi Berkelhamer of the University of California at Irvine for the data. The data are part of a larger set collected for an undergraduate lab exercise in scientific methods.

22. See Note 8 for Chapter 3.

Chapter 25

1. L. Loffredo et al., “Dark chocolate acutely improves walking autonomy in patients with peripheral artery disease,” *Journal of the American Heart Association*, 3 (2014), e001072. doi:10.1161/JAHA.114.001072.

2. M.-H. Chen et al., “Influence of relative age on diagnosis and treatment of attention-deficit hyperactivity disorder in Taiwanese children,” *Journal of Pediatrics*, 176 (2016), pp. 162–167. doi:10.1016/j.jpeds.2016.02.012.

3. I. M. Libman et al., “Effect of metformin added to insulin on glycemic control among overweight/obese adolescents with type 1 diabetes: A randomized clinical trial,” *Journal of the American Medical Association*, 314 (2015), pp. 2241–2250. doi:10.1001/jama.2015.16174.

4. E. R. Hunt et al., “Ants show a leftward turning bias when exploring unknown nest sites,” *Biology Letters*, 10 (2014), 20140945. doi:10.1098/rsbl.2014.0945.

5. Y.-Y. Fan et al., “Effects of storage temperature and duration on release of antimony and bisphenol A from polyethylene terephthalate drinking water bottles of China,” *Environmental Pollution*, 192 (2014), pp. 113–120. doi:10.1016/j.envpol.2014.05.012.

6. See Note 5 for Chapter 1.

7. C. Li et al., “25-Hydroxycholesterol protects host against Zika virus infection and its associated microcephaly in a mouse model,” *Immunity*, 46 (2017), pp. 1–11. doi:10.1016/j.immuni.2017.02.012.

8. B. G. Brady et al., “Self-reported adverse tattoo reactions: A New York City Central Park study, *Contact Dermatitis*, 73 (2015), pp. 91–99.

doi:10.1111/cod.12425.

9. E. C. Yip et al., “Coy males and seductive females in the sexually cannibalistic colonial spider, *Cyrtophora citricola*,” *PLOS ONE*, 11 (2016), e0155433. doi:10.1371/journal.pone.0155433.

10. M. K. Pawlik, “The effect of ginkgo biloba on the post-lunch dip and chemosensory function,” MS thesis, Purdue University, 2002.

11. O. Zerbo et al., “Association between influenza infection and vaccination during pregnancy and risk of autism spectrum disorder,” *JAMA Pediatrics*, 171 (2017), e163609. doi:10.1001/jamapediatrics.2016.3609.

12. The data are part of a larger data set in the *Journal of Statistics Education* archive, accessible online. The original source is P. Brofeldt, “Bidrag till kaennedom on fiskbestondet i vaara sjoear. Laengelmaevesi,” in T. H. Jaervi, *Finlands fiskeriet*, Vol. 4, *Meddelanden utgivna av fiskerifoereningen i Finland*, Helsinki, 1917. The data were contributed to the archive (with information in English) by Juha Puranen of the University of Helsinki.

13. Data provided by Corinne Lim, Purdue University, from a student project supervised by Professor Joseph Vanable.

14. See Note 2 for Chapter 16. The exercises are simplified, in that the measures reported in this paper have been statistically adjusted for “sociodemographic status.”

15. See Note 29 for Chapter 1.

16. From a graph in S. Boutin et al., “Anticipatory reproduction and population growth in seed predators,” *Science*, 314 (2006), pp. 1928–1930.

17. M. Xu and R. G. Palmer, “Genetic analysis of 4 new mutants at the unstable k2 Mdh1-n y20 chromosomal region in soybean,” *Journal of Heredity*, 97 (2006), pp. 423–427.

18. See Note 31 for Chapter 1.

19. See Note 21 for Chapter 24.

20. J. L. Warner-Schmidt et al., “Antidepressant effects of selective serotonin reuptake inhibitors (SSRIs) are attenuated by antiinflammatory drugs in mice and humans,” *Proceedings of the*

National Academy of Sciences, 108 (2011), pp. 9262–9267. doi:10.1073/pnas.1104836108.

21. See Note 15 for Chapter 8.

22. See Note 10 for Chapter 7.

23. See Note 14 for Chapter 8.

24. See Note 26 for Chapter 4.

Chapter 26

1. P. M. Johnson and P. J. Kenny, “Dopamine D2 receptors in addiction-like reward dysfunction and compulsive eating in obese rats, *Nature Neuroscience*, 13 (2010), pp. 635–641, doi:10.1038/nn.2519.

2. If your software doesn’t provide automated tests for multiple comparisons and you don’t have access to a table of Tukey critical values like Table G, you can perform the Bonferroni procedures by hand for a simple, though somewhat conservative, approximation. For any *c* multiple-comparisons tests, the Bonferroni approach uses traditional *t* tests but requires a significance level reduced by a factor of cc for each test (that is,α=overall significance level divided by c). (that is, $\alpha =$ overall significance level divided by c). For example, if you want an overall significance level of 5% and run 10 multiple-comparisons tests, then each test will be declared significant only if its *P*-value is 0.05/10=0.005$0.05/10 = 0.005$ or less.

3. A. Adan and J. M. Serra-Grabulosa, “Effects of caffeine and glucose, alone and combined, on cognitive performance,” *Human Psychopharmacology: Clinical and Experimental*, 25 (2010), pp. 310–317, doi:10.1002/hup.1115.

4. We thank Charles Cannon of Duke University for providing the data. The study report is C. H. Cannon, D. R. Peart, and M. Leighton, “Tree species diversity in commercially logged Bornean rainforest,” *Science*, 281 (1998), pp. 1366–1367.

5. Modified from M. C. Wilson and R. E. Shade, “Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow spittlebug,” *Journal of Economic Entomology*, 60 (1967), pp. 578–580.

6. R. Aroca et al., “Arbuscular mycorrhizal symbiosis influences strigolactone production under salinity and alleviates salt stress in lettuce plants,” *Journal of Plant Physiology*, 170 (2013), pp. 47–55, doi:10.1016/j.jplph.2012.08.020.

7. F. H. Simmons, “Physiology of the trade-off between fecundity and survival in *Drosophila melanogaster*, as revealed through dietary manipulation,” MS thesis, University of California, Irvine, 1996.

8. Data courtesy of David LeBauer, University of California, Irvine.

9. Simulated data, based on data summaries in G. L. Cromwell et al., “A comparison of the nutritive value of *opaque-2*, *floury-2* and normal corn for the chick,” *Poultry Science*, 57 (1968), pp. 840–847.

10. We thank Andy Niemiec and Robbie Molden for data from a summer science project at Kenyon College. Provided by Brad Hartlaub.

11. Data courtesy of Brad Hughes, University of California, Irvine.

12. Hematology data for Lead(2+) Acetate from National Toxicology Program Study No. IMM92043, http://ntp-server.niehs.nih.gov.

13. We thank Rudi Berkelhamer of the University of California at Irvine for the data. The data are part of a larger set collected for an undergraduate lab exercise in scientific methods.

14. B.Stricanne, “Etudes d’intÉgration multisensorielles dans la voie visuelle occipito-pariÉtale du primate,” PhD thesis, UniversitÉ Paris VI, 1996.

15. C.E. Mowen, “Use of glusofinate in glusofinate resistant corn hybrids,” MS thesis, Purdue University, 1999.

Chapter 27

1. Data provided by Samuel Phillips, Purdue University.

2. G. Yeretssian et al., “Is BID required for NOD signalling?” *Nature*, 474 (2011), pp. 96–99, doi:10.1038/nature11367.

3. F. H. Simmons, “Physiology of the trade-off between fecundity and survival in *Drosophila melanogaster*, as revealed through dietary manipulation,” MS thesis, University of California at Irvine, 1996.

4. The precise meaning of “yields are systematically larger in plots with no weeds” is that for every fixed value *a*, the probability that the yield with no weeds is larger than *a* is at least as great as the same probability for the yield with weeds.

5. H. Chern Boo, “Consumers’ perceptions and concerns about safety and healthfulness of food served at fairs and festivals,” MS thesis, Purdue

University, 1997.

6. From a graph in F. Grieco, A. J. van Noordwijk, and M. E. Visser, “Evidence for the effect of learning on timing of reproduction in blue tits,” *Science*, 296 (2002), pp. 136–138.

7. We thank David LeBauer of the University of California at Irvine for providing the data.

8. N.Guéeguen and C. Petr, “Odors and consumer behavior in a restaurant,” *Journal of Hospitality Management*, 25 (2006), pp. 335–339. We thank Nicolas Guéguen for providing the data.

9. B.Stricanne, “Etudes d’intégration multisensorielles dans la voie visuelle occipito-pariétale du primate,” PhD thesis, Université Paris VI, 1996.

10. R. Sakakibara et al., “Influence of body position on defecation in humans,” *Lower Urinary Tract Symptoms*, 2 (2010), pp. 16–21, doi:10.1111/j.1757-5672.2009.00057.x.

11. R. A. Berner and G. P. Landis, “Gas bubbles in fossil amber as possible indicators of the major gas composition of ancient air,” *Science*, 239 (1988), pp. 1406–1409.

12. A.Fernández-Nebro et al., “Treatment of rheumatic inflammatory disease in 25 patients with secondary amyloidosis using tumor necrosis factor alpha antagonists,” *American Journal of Medicine*, 118 (2005), pp. 552–556.

13. See Note 1.

14. Modified from M. C. Wilson and R. E. Shade, “Relative attractiveness ofvarious Luminescent colors to the cereal leaf beetle and the meadow spittlebug,” *Journal of Economic Entomology*, 60 (1967), pp. 578–580.

15. Data provided by Warren Page, New York City Technical College, from a study done by John Hudesman.

16. R. A. Morgan et al., “Cancer regression in patients after transfer of genetically engineered lymphocytes,” *Science*, 314 (2006), pp. 126–129. The data appear in the Online Supplementary Material.

17. N. Ogata et al., “Brain structural abnormalities in Doberman pinschers with canine compulsive disorder,” *Progress in Neuro-psychopharmacology and Biological Psychiatry*, 45 (2013), pp. 1–6, doi:10.1016/j.pnpbp.2013.04.002.

18. Data provided by Matthew Moore.

19. We thank Brad Hughes of the University of California at Irvine for providing the data.

20. D. H. Barouch et al., “Vaccine protection against acquisition of neutralization- resistant SIV challenges in rhesus monkeys,” *Nature,* 482 (2012), pp. 89–93, doi:10.1038/nature10766.

21. L. H. Yang, “Periodical cicadas as resource pulses in North American forests,” *Science,* 306 (2004), pp. 1565–1567. The data are simulated Normal values that match the means and standard deviations reported in this article.

22. N. A. Sherif et al., “Detection of cotinine in neonate meconium as a marker for nicotine exposure in utero,” *Eastern Mediterranean Health Journal,* 10 (2004), pp. 96–105.

23. Y.-H. Huang et al., “Nanoparticle-delivered suicide gene therapy effectively reduces ovarian tumor burden in mice,” *Cancer Research,* 69 (2009), pp. 6184–6191.

24. *Black Newsprint Inks,* National Toxicology Program Toxicity Report Series No. 17, ntp.niehs.nih.gov.

25. C. Cox Fernandes, J. Podos, and J. G. Lundberg, “Amazonian ecology: tributaries enhance the diversity of electric fishes,” *Science,* 305 (2004), pp. 1960–1962.

Chapter 28

1. Data were estimated from a scatterplot in L. Partridge and M. Farquhar, “Sexual activity reduces lifespan of male fruitflies,” *Nature,* 294 (1981), pp. 580–582.

2. Data were estimated from a scatterplot in T. J. Cleophas, “The sense and nonsense of regression modeling for increasing precision of clinical trials,” *Clinical Pharmacology and Therapeutics*, 74 (2003), pp. 295–297.

3. Data were estimated from a scatterplot in P. Heeb, M. Kolliker, and H. Richner, “Bird-ectoparasite interactions, nest humidity, and ectoparasite community structure,” *Ecology*, 81 (2000), pp. 958–968.

4. F. H. Simmons, “Physiology of the trade-off between fecundity and survival in *Drosophila melanogaster*, as revealed through dietary manipulation,” MS thesis, University of California at Irvine, 1996.

5. Data courtesy of Brad Hughes, Department of Ecology and Evolutionary Biology, University of California, Irvine.

6. M. Di Monacoa et al., “Biochemical markers of nutrition and bone mineral density in the elderly,” *Gerontology*, 49 (2003), pp. 50–54.

7. S. J. Husson et al., “Optogenetic analysis of a nociceptor neuron and network reveals ion channels acting downstream of primary sensors,” *Current Biology*, 22 (2012), pp. 743–752, doi:10.1016/j.cub.2012.02.066.

8. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by M. J. Kahn of Wheaton College.

9. D. P. Casey et al., “Relationship between muscle sympathetic nerve activity and aortic wave reflection characteristics in young men and women*,” Hypertension*, 57 (2011), pp. 421–427.

10. R. Margaria et al., “Energy cost of running,” *Journal of Applied Physiology*, 18 (1963), pp. 367–370.

11. The data in Table 28.4 are part of a larger data set in the *Journal of Statistics Education* online archive. The original source is P. Brofeldt, “Bidrag till kaennedom on fiskbestondet i vaara sjoear. Laengelmaevesi,” in T. H. Jaervi, *Finlands fiskeriet*, vol. 4, *Meddelanden utgivna av fiskerifoereningen i Finland*, Helsinki, 1917. The data were contributed to the archive (with information in English) by J. Puranen of the University of Helsinki.

12. These data come from the Data and Story Library online at lib.stat.cmu.edu/DASL. They were contributed by D. S. Moore and G. P. McCabe.

13. A.Spanos, F. E. Harrell, and D. T. Durack, “Differential diagnosis of acute meningitis: an analysis of the predictive value of initial observations,” *Journal of the American Medical Association*, 262 (1989), pp. 2700–2707.

14. These data come from the Data and Story Library online at lib.stat.cmu.edu/DASL and were first published in L. J. Wei, D. Y.

Lin, and L. Weissfeld, “Tumor recurrence data for patients with bladder cancer,” *Journal of the American Statistical Association*, 84 (1989), pp. 1065–1073.

15. The test statistic for logistic regression is *z*, not *t* (unlike for simple and multiple linear regression). Interestingly, when *z* follows the standard Normal distribution, the value z2 z^2 actually follows a chi-square distribution with df = 1. For this reason, inference forlogistic regression sometimes uses the chi-square statistic (such that X2=z2 $X^2 = z^2$) and the chi-square distribution with df = 1. Both methods give the same *P*-value.

16. E. de Jonge et al., “Effects of selective decontamination of digestive tract on mortality and acquisition of resistant bacteria in intensive care: a randomised controlled trial,” *Lancet*, 362 (2003), pp. 1011–1016.

17. A. M. Tromp et al., “Predictors for falls and fractures in the longitudinal aging study, Amsterdam,” *Journal of Bone and Mineral Research*, 13 (1998), pp. 1932–1939.

18. This data set comes from the *Journal of Statistics Education* online data archive. It was originally submitted by M. C. Meyer from the University of Georgia.

19. We thank Haruhiko Itagaki and his students Andrew Vreede and Marissa Stearns for providing data on tobacco hornworm caterpillars (*Manduca sexta*).

20. Data courtesy of David LeBauer, University of California at Irvine.

21. P. Velleman, *ActivStats 2.0,* Addison-Wesley Interactive, 1997.

22. J. T. Fleming, “The measurement of children’s perception of difficulty in reading materials,” *Research in the Teaching of English*, 1 (1967), pp. 136–156.

23. For more details, see H. Hoppeler and E. Weibel, “Scaling functions to body size: theories and facts,” *Journal of Experimental Biology*, 208 (2005), pp. 1573–1574.

24. J. B. Schorling et al., “A trial of church-based smoking cessation interventions for rural African Americans,” *Preventive Medicine,* 26 (1997), pp. 92–101. Data posted on the website of the Department of Biostatistics at Vanderbilt University.

25. D. Gordon, “The relationships among academic self-concept, academic achievement, and persistence with academic self-attribution, study habits, and perceived school environment,” PhD thesis, Purdue University, 1997.

26. We thank Ray and Pat Heithaus for providing data on the pine seedlings at the Brown Family Environmental Center.

Notes

Brief Contents

[i] **Starred material is optional.**

Contents

[i] **Starred material is optional.**

To the Instructor: About this Book

[i] *SPSS was acquired by IBM in October 2009.*

Chapter 1

[i] *You can visualize this and many other data sets directly on the California open data portal.*

Chapter 2

[i] *This section is optional.*

Chapter 3

[i] *Optional Chapter 28 on multiple regression provides more detailed coverage of this topic.*

Chapter 4

[i] *This section is optional.*

Chapter 9

[i] *This section is optional.*

Chapter 10

[i] *This chapter gives more details about probability that are not needed to read the rest of the book. These concepts, however, are particularly important in medicine and clinical research.*

Chapter 11

[i] *This section is unnecessary if you will always use software for Normal distribution calculations.*

[ii] *This section is optional for a quarter-based course.*

Chapter 12

[i] *This more advanced chapter concerns a special topic in probability. It is not needed to read the rest of the book.*

Chapter 13

[i] *This section is optional.*

Chapter 15

[i] *The remainder of this chapter presents more advanced material that is not needed to read the rest of the book. The idea of planning studies is, however, very important in practice.*

Chapter 17

[i] *This computational section is optional.*

Chapter 18

[i] *These two short sections concern optional special topics. They are not needed to read the rest of the book but can serve as background for Chapter 24 on one-way ANOVA.*

Chapter 19

[i] *This section presents more advanced material that is not needed to read the rest of the book. The idea of planning studies is, however, very important in practice.*

Chapter 20

[i] *The remainder of this chapter presents more advanced material that is not needed to read the rest of the book. Relative risk and odds ratio are, however, important concepts in the health sciences.*

Chapter 22

[i] *This section is optional.*

Chapter 23

[i] *This section is optional.*

Chapter 24

[i] *This more advanced section is optional if you are using software to find the F statistic.*

Chapter 26

[i] *This material is optional.*

[i] *This optional material requires the optional section on some details of ANOVA from Chapter 24, page 621.*

TABLES

Table A: Random Digits

Table B: Standard Normal Probabilities

Table C: *t* Distribution Critical Values

Table D: Chi-square Distribution Critical Values

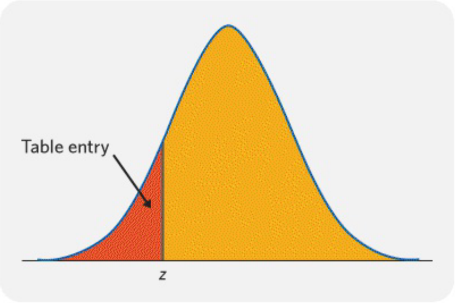
Table E: Critical Values of the Correlation *r*

Table F: *F* Distribution Critical Values

Table G: Critical values m* for Tukey pairwise multiple comparisons with 95% confidence level, *k* comparisons, and *N* – *k* degrees of freedom (df)

TABLE A: Random Digits

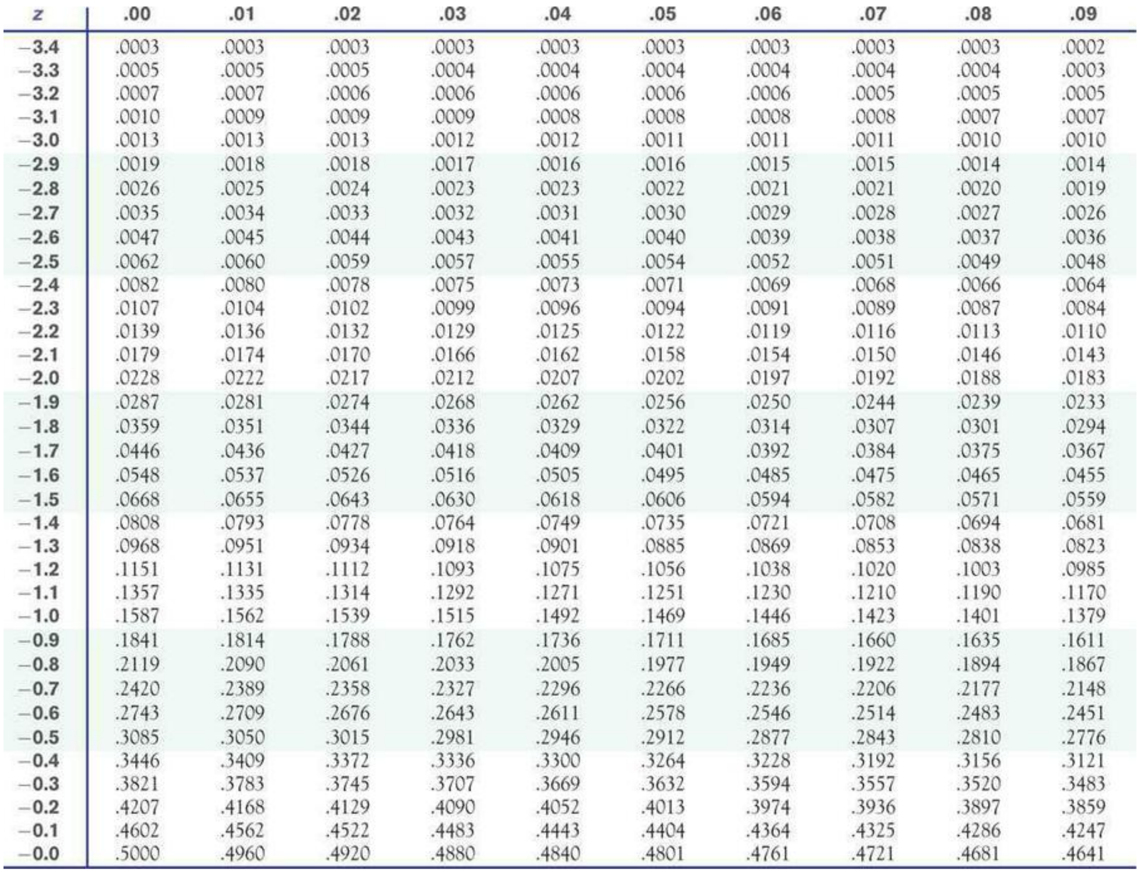
Line								
101	19223	95034	05756	28713	96409	12531	42544	82853
102	73676	47150	99400	01927	27754	42648	82425	36290
103	45467	71709	77558	00095	32863	29485	82226	90056
104	52711	38889	93074	60227	40011	85848	48767	52573
105	95592	94007	69971	91481	60779	53791	17297	59335
106	68417	35013	15529	72765	85089	57067	50211	47487
107	82739	57890	20807	47511	81676	55300	94383	14893
108	60940	72024	17868	24943	61790	90656	87964	18883
109	36009	19365	15412	39638	85453	46816	83485	41979
110	38448	48789	18338	24697	39364	42006	76688	08708
111	81486	69487	60513	09297	00412	71238	27649	39950
112	59636	88804	04634	71197	19352	73089	84898	45785
113	62568	70206	40325	03699	71080	22553	11486	11776
114	45149	32992	75730	66280	03819	56202	02938	70915
115	61041	77684	94322	24709	73698	14526	31893	32592
116	14459	26056	31424	80371	65103	62253	50490	61181
117	38167	98532	62183	70632	23417	26185	41448	75532
118	73190	32533	04470	29669	84407	90785	65956	86382
119	95857	07118	87664	92099	58806	66979	98624	84826
120	35476	55972	39421	65850	04266	35435	43742	11937
121	71487	09984	29077	14863	61683	47052	62224	51025
122	13873	81598	95052	90908	73592	75186	87136	95761
123	54580	81507	27102	56027	55892	33063	41842	81868
124	71035	09001	43367	49497	72719	96758	27611	91596
125	96746	12149	37823	71868	18442	35119	62103	39244
126	96927	19931	36809	74192	77567	88741	48409	41903
127	43909	99477	25330	64359	40085	16925	85117	36071
128	15689	14227	06565	14374	13352	49367	81982	87209
129	36759	58984	68288	22913	18638	54303	00795	08727
130	69051	64817	87174	09517	84534	06489	87201	97245
131	05007	16632	81194	14873	04197	85576	45195	96565
132	68732	55259	84292	08796	43165	93739	31685	97150
133	45740	41807	65561	33302	07051	93623	18132	09547
134	27816	78416	18329	21337	35213	37741	04312	68508
135	66925	55658	39100	78458	11206	19876	87151	31260
136	08421	44753	77377	28744	75592	08563	79140	92454
137	53645	66812	61421	47836	12609	15373	98481	14592
138	66831	68908	40772	21558	47781	33586	79177	06928
139	55588	99404	70708	41098	43563	56934	48394	51719
140	12975	13258	13048	45144	72321	81940	00360	02428
141	96767	35964	23822	96012	94591	65194	50842	53372
142	72829	50232	97892	63408	77919	44575	24870	04178
143	88565	42628	17797	49376	61762	16953	88604	12724
144	62964	88145	83083	69453	46109	59505	69680	00900
145	19687	12633	57857	95806	09931	02150	43163	58636
146	37609	59057	66967	83401	60705	02384	90597	93600
147	54973	86278	88737	74351	47500	84552	19909	67181
148	00694	05977	19664	65441	20903	62371	22725	53340
149	71546	05233	53946	68743	72460	27601	45403	88692
150	07511	88915	41267	16853	84569	79367	32337	03316



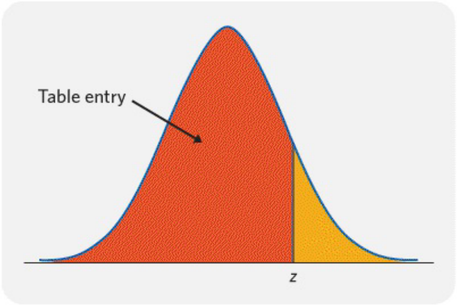
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

Table entry for *z* is the area under the standard Normal curve to the left of *z*.

TABLE B: Standard Normal Probabilities

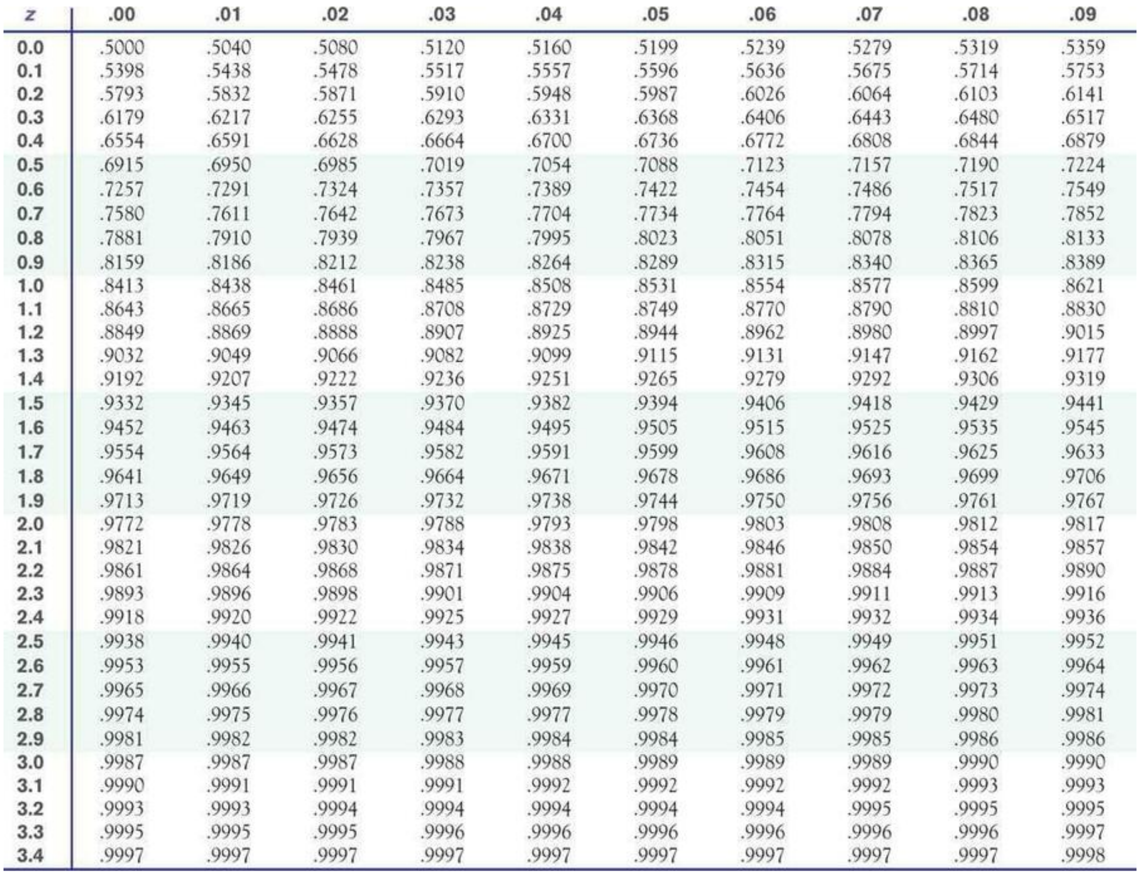


z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

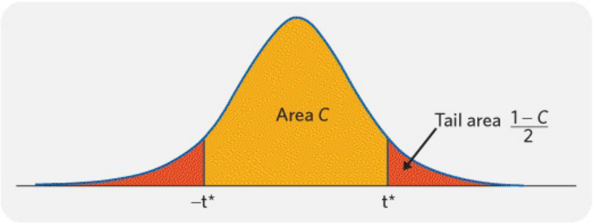


Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

Table entry for *z* is the area under the standard Normal curve to the left of *z*.



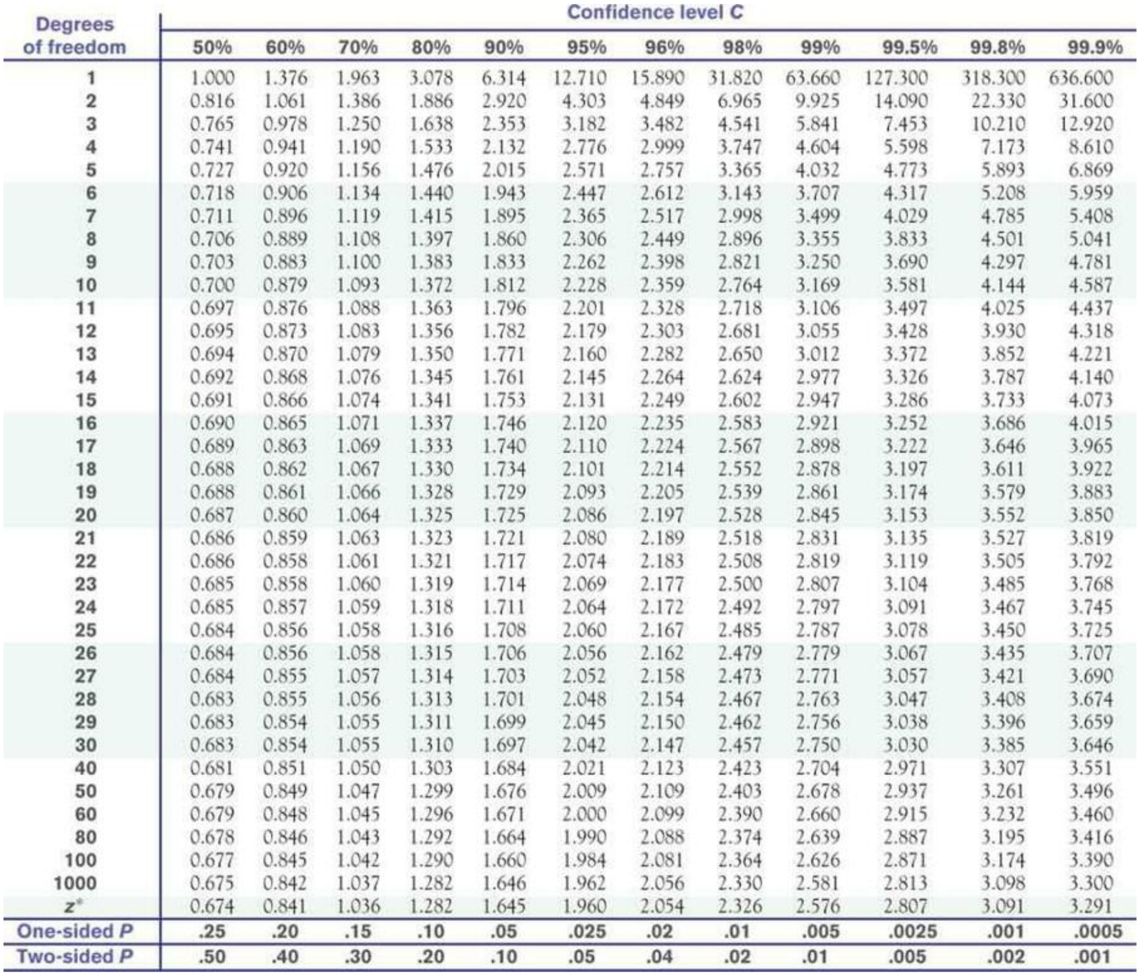
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998



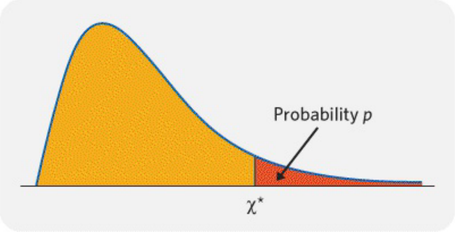
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

Table entry for *C* is the critical value t^* required for confidence level *C*. To approximate one- and two-sided *P*-values, compare the value of the *t* statistic with the critical values of t^* that match the *P*-values given at the bottom of the table.

TABLE C: *t* Distribution Critical Values



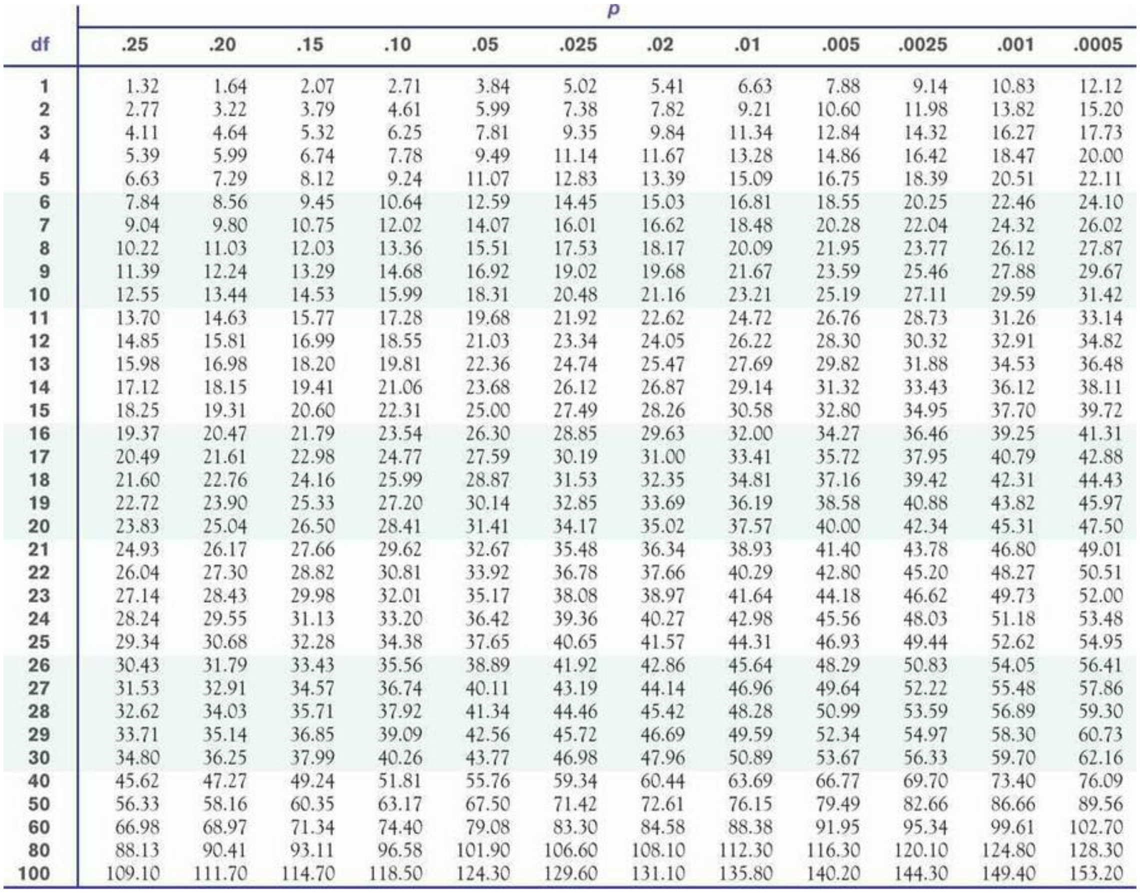
Degrees of freedom	Confidence level C											
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
1	1.000	1.376	1.963	3.078	6.314	12.710	15.890	31.820	63.660	127.300	318.300	636.600
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.090	22.330	31.600
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.210	12.920
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
One-sided P	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
Two-sided P	.50	.40	.30	.20	.10	.05	.04	.02	.01	.005	.002	.001



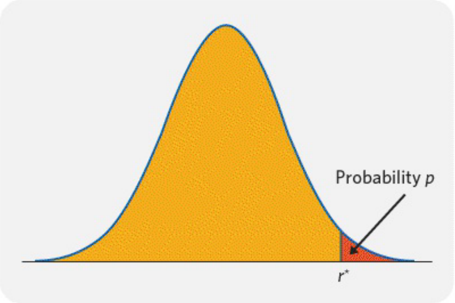
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

Table entry for *p* is the critical value *X** with probability *p* lying to its right.

TABLE D: Chi-square Distribution Critical Values



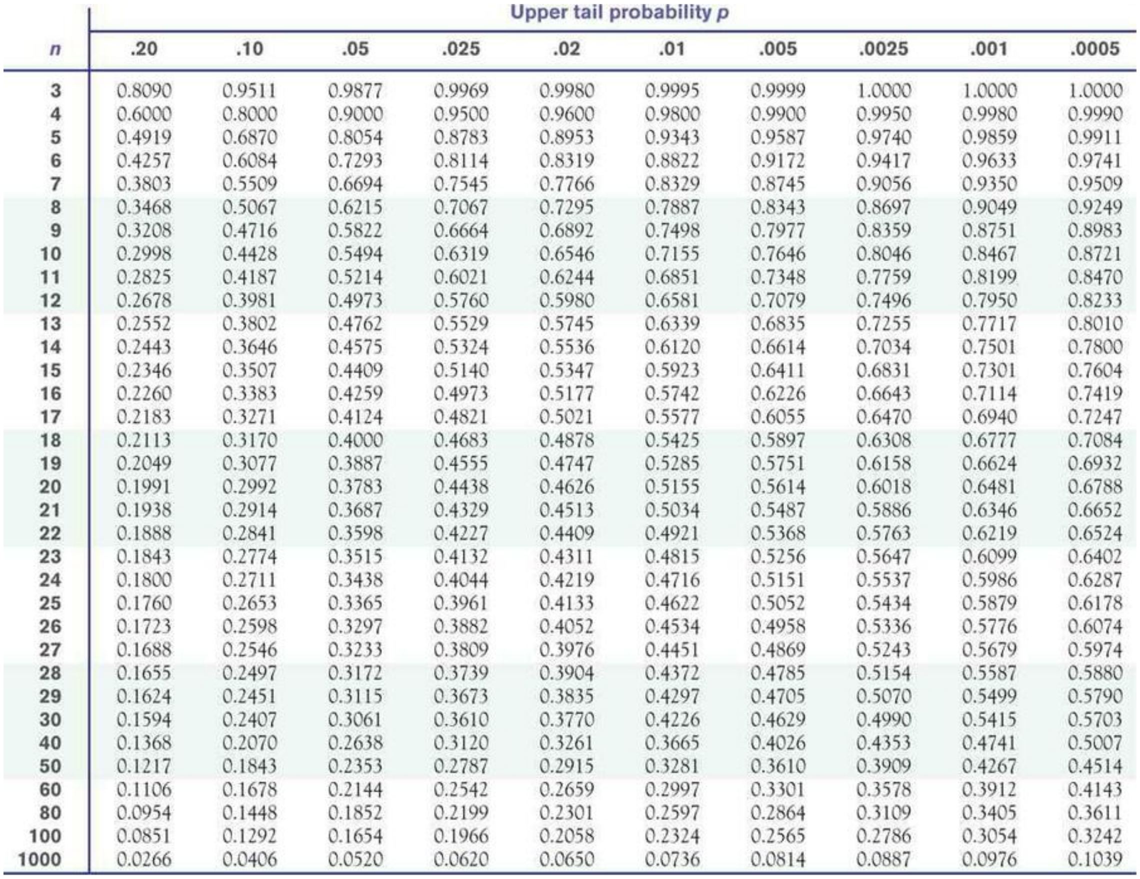
	p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.70
80	88.13	90.41	93.11	96.58	101.90	106.60	108.10	112.30	116.30	120.10	124.80	128.30
100	109.10	111.70	114.70	118.50	124.30	129.60	131.10	135.80	140.20	144.30	149.40	153.20



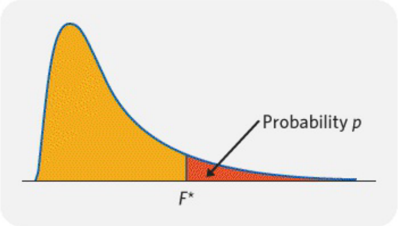
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e,
© 2018 W. H. Freeman and Company

Table entry for *p* is the critical value *r** of the correlation coefficient *r* with probability *p* lying to its right.

TABLE E: Critical Values of the Correlation *r*



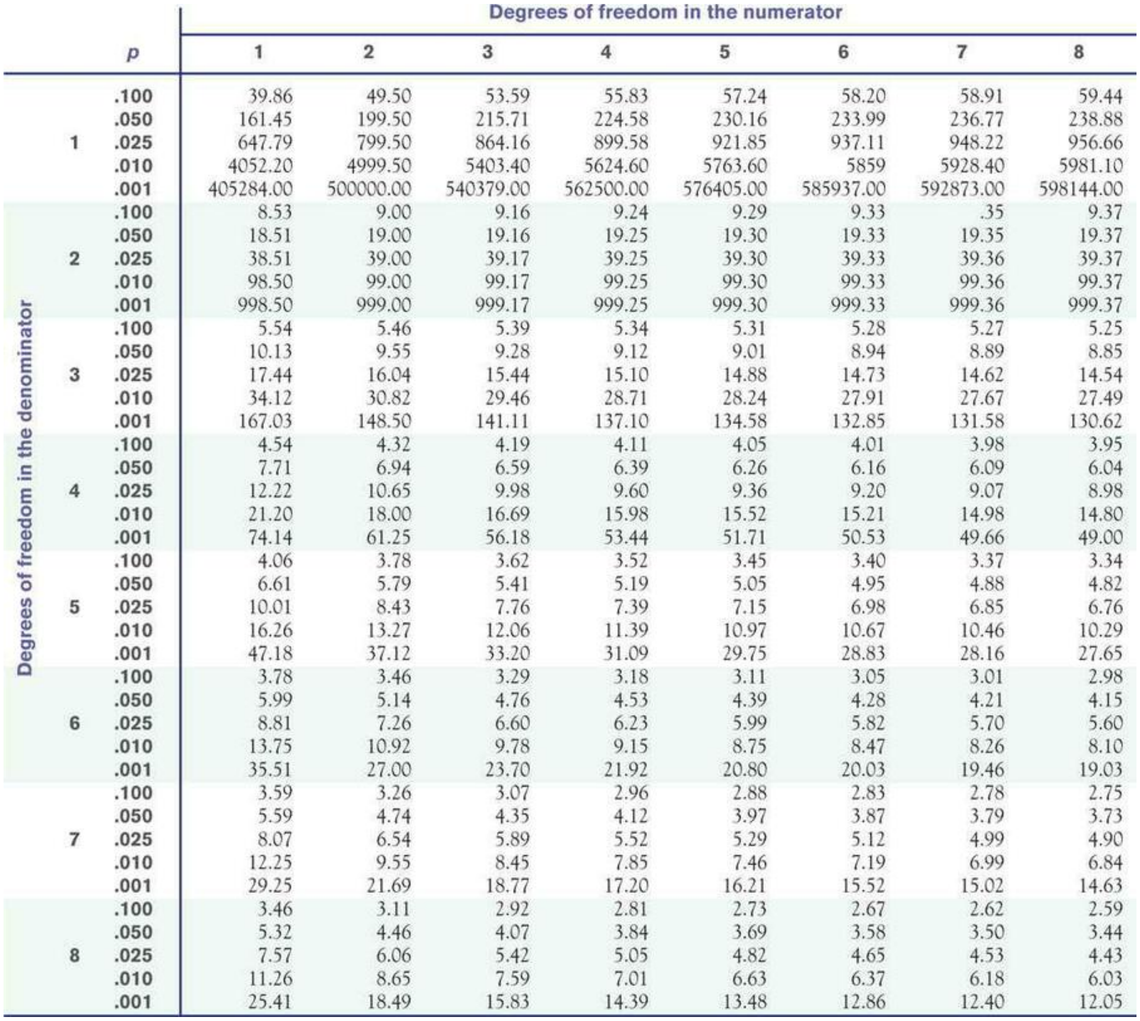
	Upper tail probability *p*									
n	.20	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
3	0.8090	0.9511	0.9877	0.9969	0.9980	0.9995	0.9999	1.0000	1.0000	1.0000
4	0.6000	0.8000	0.9000	0.9500	0.9600	0.9800	0.9900	0.9950	0.9980	0.9990
5	0.4919	0.6870	0.8054	0.8783	0.8953	0.9343	0.9587	0.9740	0.9859	0.9911
6	0.4257	0.6084	0.7293	0.8114	0.8319	0.8822	0.9172	0.9417	0.9633	0.9741
7	0.3803	0.5509	0.6694	0.7545	0.7766	0.8329	0.8745	0.9056	0.9350	0.9509
8	0.3468	0.5067	0.6215	0.7067	0.7295	0.7887	0.8343	0.8697	0.9049	0.9249
9	0.3208	0.4716	0.5822	0.6664	0.6892	0.7498	0.7977	0.8359	0.8751	0.8983
10	0.2998	0.4428	0.5494	0.6319	0.6546	0.7155	0.7646	0.8046	0.8467	0.8721
11	0.2825	0.4187	0.5214	0.6021	0.6244	0.6851	0.7348	0.7759	0.8199	0.8470
12	0.2678	0.3981	0.4973	0.5760	0.5980	0.6581	0.7079	0.7496	0.7950	0.8233
13	0.2552	0.3802	0.4762	0.5529	0.5745	0.6339	0.6835	0.7255	0.7717	0.8010
14	0.2443	0.3646	0.4575	0.5324	0.5536	0.6120	0.6614	0.7034	0.7501	0.7800
15	0.2346	0.3507	0.4409	0.5140	0.5347	0.5923	0.6411	0.6831	0.7301	0.7604
16	0.2260	0.3383	0.4259	0.4973	0.5177	0.5742	0.6226	0.6643	0.7114	0.7419
17	0.2183	0.3271	0.4124	0.4821	0.5021	0.5577	0.6055	0.6470	0.6940	0.7247
18	0.2113	0.3170	0.4000	0.4683	0.4878	0.5425	0.5897	0.6308	0.6777	0.7084
19	0.2049	0.3077	0.3887	0.4555	0.4747	0.5285	0.5751	0.6158	0.6624	0.6932
20	0.1991	0.2992	0.3783	0.4438	0.4626	0.5155	0.5614	0.6018	0.6481	0.6788
21	0.1938	0.2914	0.3687	0.4329	0.4513	0.5034	0.5487	0.5886	0.6346	0.6652
22	0.1888	0.2841	0.3598	0.4227	0.4409	0.4921	0.5368	0.5763	0.6219	0.6524
23	0.1843	0.2774	0.3515	0.4132	0.4311	0.4815	0.5256	0.5647	0.6099	0.6402
24	0.1800	0.2711	0.3438	0.4044	0.4219	0.4716	0.5151	0.5537	0.5986	0.6287
25	0.1760	0.2653	0.3365	0.3961	0.4133	0.4622	0.5052	0.5434	0.5879	0.6178
26	0.1723	0.2598	0.3297	0.3882	0.4052	0.4534	0.4958	0.5336	0.5776	0.6074
27	0.1688	0.2546	0.3233	0.3809	0.3976	0.4451	0.4869	0.5243	0.5679	0.5974
28	0.1655	0.2497	0.3172	0.3739	0.3904	0.4372	0.4785	0.5154	0.5587	0.5880
29	0.1624	0.2451	0.3115	0.3673	0.3835	0.4297	0.4705	0.5070	0.5499	0.5790
30	0.1594	0.2407	0.3061	0.3610	0.3770	0.4226	0.4629	0.4990	0.5415	0.5703
40	0.1368	0.2070	0.2638	0.3120	0.3261	0.3665	0.4026	0.4353	0.4741	0.5007
50	0.1217	0.1843	0.2353	0.2787	0.2915	0.3281	0.3610	0.3909	0.4267	0.4514
60	0.1106	0.1678	0.2144	0.2542	0.2659	0.2997	0.3301	0.3578	0.3912	0.4143
80	0.0954	0.1448	0.1852	0.2199	0.2301	0.2597	0.2864	0.3109	0.3405	0.3611
100	0.0851	0.1292	0.1654	0.1966	0.2058	0.2324	0.2565	0.2786	0.3054	0.3242
1000	0.0266	0.0406	0.0520	0.0620	0.0650	0.0736	0.0814	0.0887	0.0976	0.1039



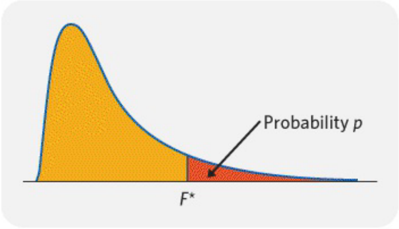
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Table entry for *p* is the critical value *r** of the correlation coefficient *r* with probability *p* lying to its right.

TABLE F: *F* Distribution Critical Values

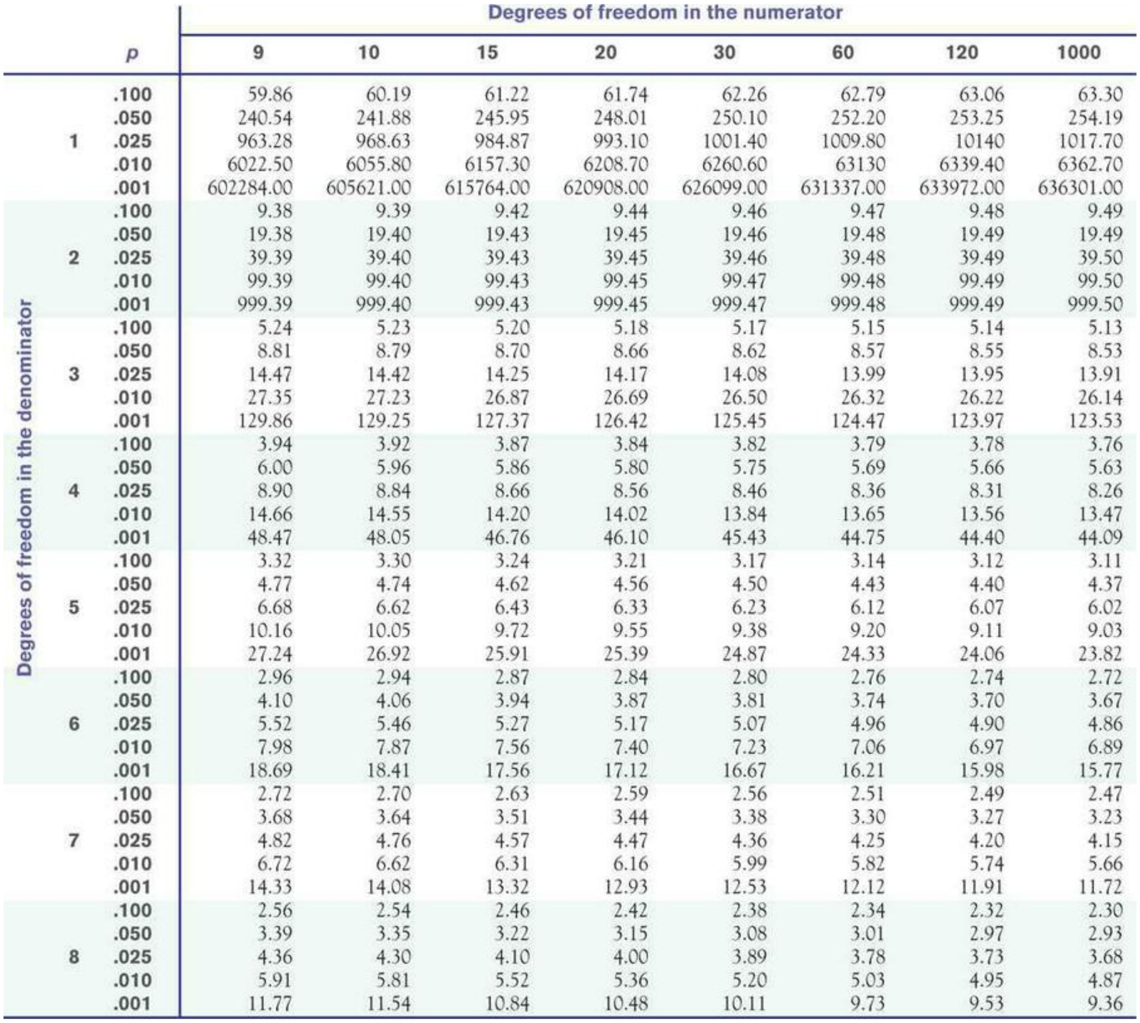


		Degrees of freedom in the numerator							
Degrees of freedom in the denominator	p	1	2	3	4	5	6	7	8
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44
	.050	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88
	.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66
	.010	4052.20	4999.50	5403.40	5624.60	5763.60	5859	5928.40	5981.10
	.001	405284.00	500000.00	540379.00	562500.00	576405.00	585937.00	592873.00	598144.00
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	.35	9.37
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37
	.001	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49
	.001	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80
	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29
	.001	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10
	.001	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84
	.001	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03
	.001	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05

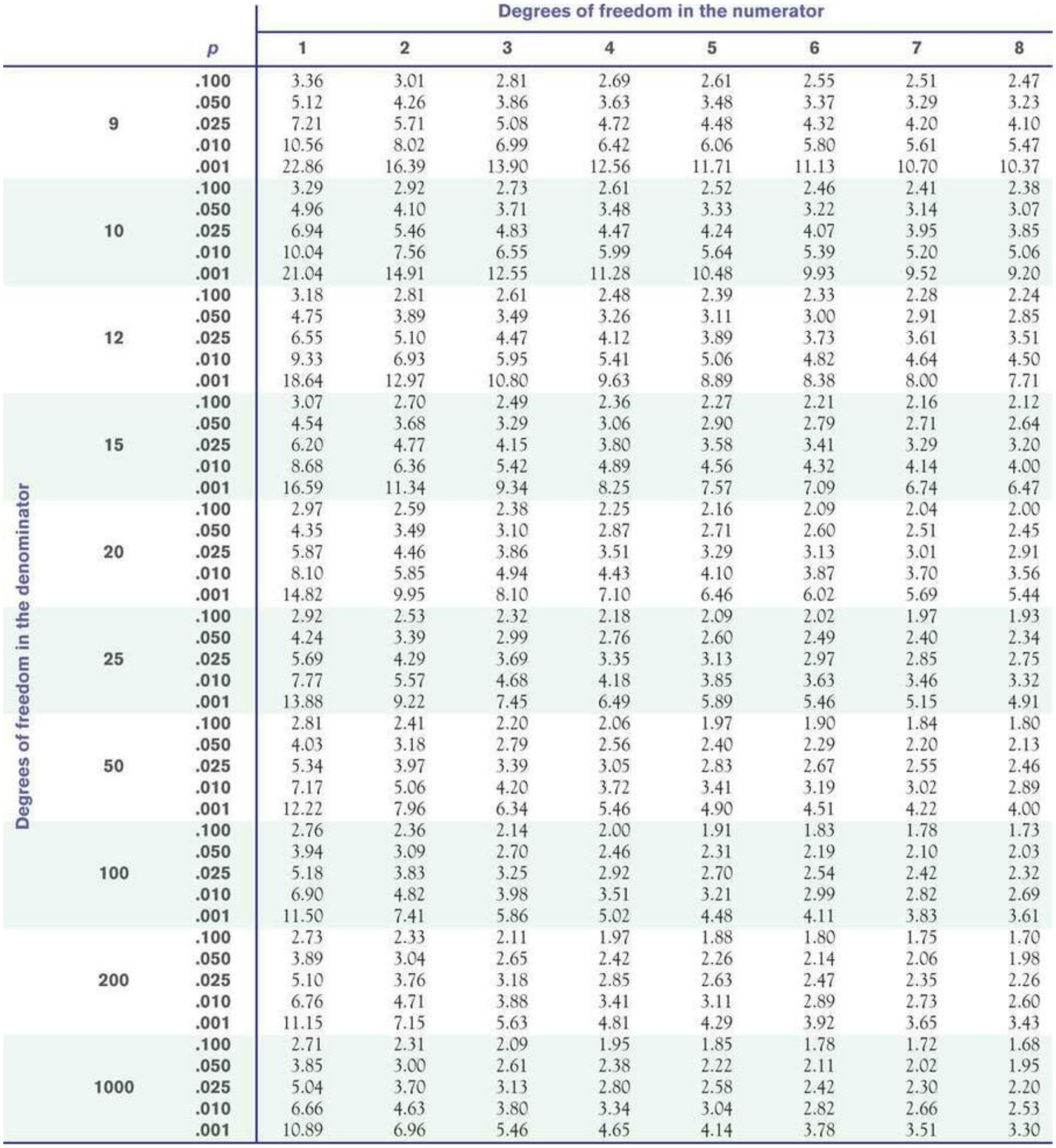


Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

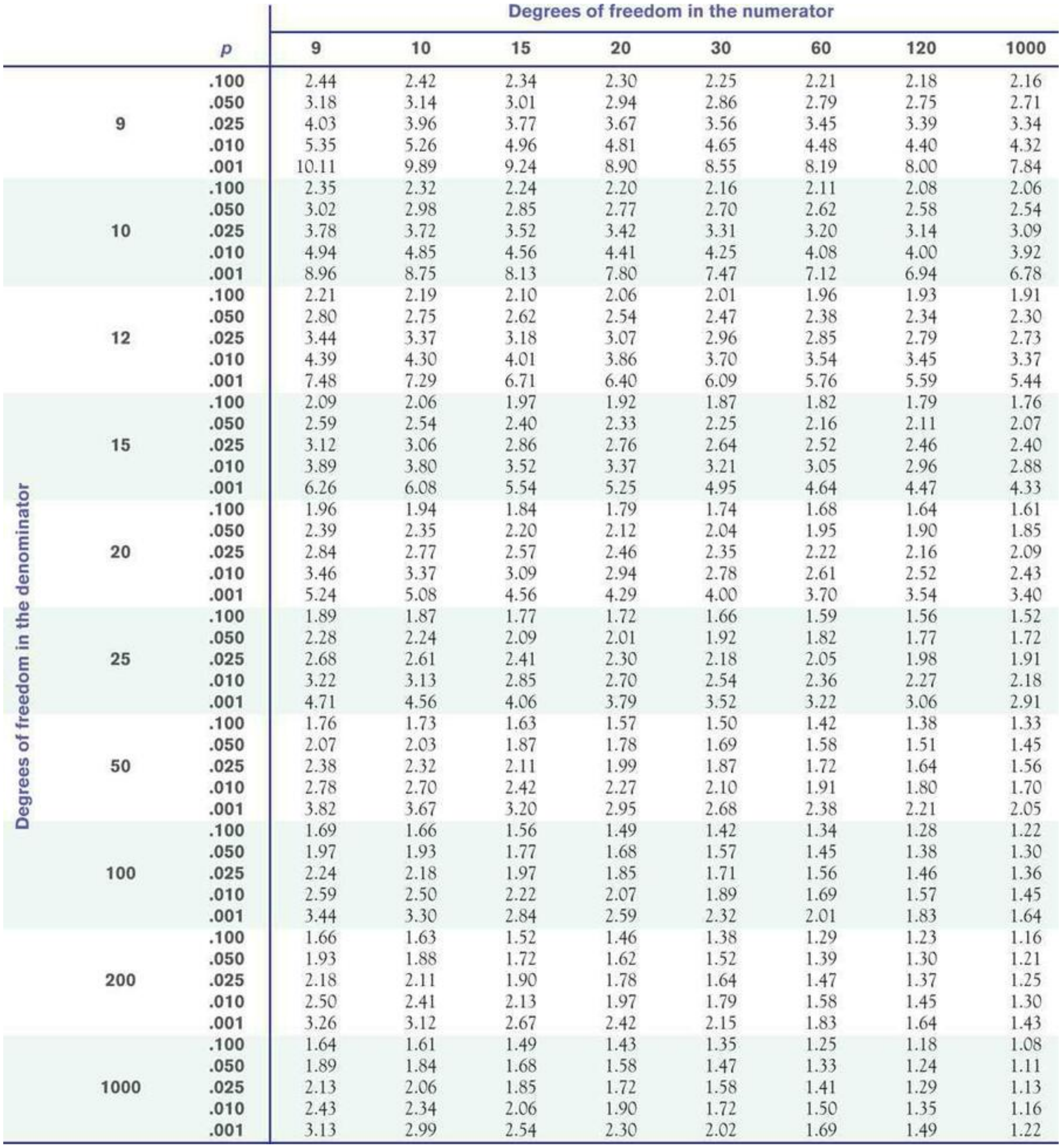
Table entry for *p* is the critical value *r** of the correlation coefficient *r* with probability *p* lying to its right.



		Degrees of freedom in the numerator							
Degrees of freedom in the denominator	**p**	**9**	**10**	**15**	**20**	**30**	**60**	**120**	**1000**
1	**.100**	59.86	60.19	61.22	61.74	62.26	62.79	63.06	63.30
	.050	240.54	241.88	245.95	248.01	250.10	252.20	253.25	254.19
	.025	963.28	968.63	984.87	993.10	1001.40	1009.80	10140	1017.70
	.010	6022.50	6055.80	6157.30	6208.70	6260.60	63130	6339.40	6362.70
	.001	602284.00	605621.00	615764.00	620908.00	626099.00	631337.00	633972.00	636301.00
2	**.100**	9.38	9.39	9.42	9.44	9.46	9.47	9.48	9.49
	.050	19.38	19.40	19.43	19.45	19.46	19.48	19.49	19.49
	.025	39.39	39.40	39.43	39.45	39.46	39.48	39.49	39.50
	.010	99.39	99.40	99.43	99.45	99.47	99.48	99.49	99.50
	.001	999.39	999.40	999.43	999.45	999.47	999.48	999.49	999.50
3	**.100**	5.24	5.23	5.20	5.18	5.17	5.15	5.14	5.13
	.050	8.81	8.79	8.70	8.66	8.62	8.57	8.55	8.53
	.025	14.47	14.42	14.25	14.17	14.08	13.99	13.95	13.91
	.010	27.35	27.23	26.87	26.69	26.50	26.32	26.22	26.14
	.001	129.86	129.25	127.37	126.42	125.45	124.47	123.97	123.53
4	**.100**	3.94	3.92	3.87	3.84	3.82	3.79	3.78	3.76
	.050	6.00	5.96	5.86	5.80	5.75	5.69	5.66	5.63
	.025	8.90	8.84	8.66	8.56	8.46	8.36	8.31	8.26
	.010	14.66	14.55	14.20	14.02	13.84	13.65	13.56	13.47
	.001	48.47	48.05	46.76	46.10	45.43	44.75	44.40	44.09
5	**.100**	3.32	3.30	3.24	3.21	3.17	3.14	3.12	3.11
	.050	4.77	4.74	4.62	4.56	4.50	4.43	4.40	4.37
	.025	6.68	6.62	6.43	6.33	6.23	6.12	6.07	6.02
	.010	10.16	10.05	9.72	9.55	9.38	9.20	9.11	9.03
	.001	27.24	26.92	25.91	25.39	24.87	24.33	24.06	23.82
6	**.100**	2.96	2.94	2.87	2.84	2.80	2.76	2.74	2.72
	.050	4.10	4.06	3.94	3.87	3.81	3.74	3.70	3.67
	.025	5.52	5.46	5.27	5.17	5.07	4.96	4.90	4.86
	.010	7.98	7.87	7.56	7.40	7.23	7.06	6.97	6.89
	.001	18.69	18.41	17.56	17.12	16.67	16.21	15.98	15.77
7	**.100**	2.72	2.70	2.63	2.59	2.56	2.51	2.49	2.47
	.050	3.68	3.64	3.51	3.44	3.38	3.30	3.27	3.23
	.025	4.82	4.76	4.57	4.47	4.36	4.25	4.20	4.15
	.010	6.72	6.62	6.31	6.16	5.99	5.82	5.74	5.66
	.001	14.33	14.08	13.32	12.93	12.53	12.12	11.91	11.72
8	**.100**	2.56	2.54	2.46	2.42	2.38	2.34	2.32	2.30
	.050	3.39	3.35	3.22	3.15	3.08	3.01	2.97	2.93
	.025	4.36	4.30	4.10	4.00	3.89	3.78	3.73	3.68
	.010	5.91	5.81	5.52	5.36	5.20	5.03	4.95	4.87
	.001	11.77	11.54	10.84	10.48	10.11	9.73	9.53	9.36



		Degrees of freedom in the numerator							
Degrees of freedom in the denominator	*p*	1	2	3	4	5	6	7	8
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47
	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06
	.001	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50
	.001	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00
	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56
	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32
	.001	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91
50	.100	2.81	2.41	2.20	2.06	1.97	1.90	1.84	1.80
	.050	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13
	.025	5.34	3.97	3.39	3.05	2.83	2.67	2.55	2.46
	.010	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89
	.001	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00
100	.100	2.76	2.36	2.14	2.00	1.91	1.83	1.78	1.73
	.050	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03
	.025	5.18	3.83	3.25	2.92	2.70	2.54	2.42	2.32
	.010	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69
	.001	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61
200	.100	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70
	.050	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98
	.025	5.10	3.76	3.18	2.85	2.63	2.47	2.35	2.26
	.010	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60
	.001	11.15	7.15	5.63	4.81	4.29	3.92	3.65	3.43
1000	.100	2.71	2.31	2.09	1.95	1.85	1.78	1.72	1.68
	.050	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95
	.025	5.04	3.70	3.13	2.80	2.58	2.42	2.30	2.20
	.010	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53
	.001	10.89	6.96	5.46	4.65	4.14	3.78	3.51	3.30



Degrees of freedom in the denominator	p	Degrees of freedom in the numerator 9	10	15	20	30	60	120	1000
9	.100	2.44	2.42	2.34	2.30	2.25	2.21	2.18	2.16
	.050	3.18	3.14	3.01	2.94	2.86	2.79	2.75	2.71
	.025	4.03	3.96	3.77	3.67	3.56	3.45	3.39	3.34
	.010	5.35	5.26	4.96	4.81	4.65	4.48	4.40	4.32
	.001	10.11	9.89	9.24	8.90	8.55	8.19	8.00	7.84
10	.100	2.35	2.32	2.24	2.20	2.16	2.11	2.08	2.06
	.050	3.02	2.98	2.85	2.77	2.70	2.62	2.58	2.54
	.025	3.78	3.72	3.52	3.42	3.31	3.20	3.14	3.09
	.010	4.94	4.85	4.56	4.41	4.25	4.08	4.00	3.92
	.001	8.96	8.75	8.13	7.80	7.47	7.12	6.94	6.78
12	.100	2.21	2.19	2.10	2.06	2.01	1.96	1.93	1.91
	.050	2.80	2.75	2.62	2.54	2.47	2.38	2.34	2.30
	.025	3.44	3.37	3.18	3.07	2.96	2.85	2.79	2.73
	.010	4.39	4.30	4.01	3.86	3.70	3.54	3.45	3.37
	.001	7.48	7.29	6.71	6.40	6.09	5.76	5.59	5.44
15	.100	2.09	2.06	1.97	1.92	1.87	1.82	1.79	1.76
	.050	2.59	2.54	2.40	2.33	2.25	2.16	2.11	2.07
	.025	3.12	3.06	2.86	2.76	2.64	2.52	2.46	2.40
	.010	3.89	3.80	3.52	3.37	3.21	3.05	2.96	2.88
	.001	6.26	6.08	5.54	5.25	4.95	4.64	4.47	4.33
20	.100	1.96	1.94	1.84	1.79	1.74	1.68	1.64	1.61
	.050	2.39	2.35	2.20	2.12	2.04	1.95	1.90	1.85
	.025	2.84	2.77	2.57	2.46	2.35	2.22	2.16	2.09
	.010	3.46	3.37	3.09	2.94	2.78	2.61	2.52	2.43
	.001	5.24	5.08	4.56	4.29	4.00	3.70	3.54	3.40
25	.100	1.89	1.87	1.77	1.72	1.66	1.59	1.56	1.52
	.050	2.28	2.24	2.09	2.01	1.92	1.82	1.77	1.72
	.025	2.68	2.61	2.41	2.30	2.18	2.05	1.98	1.91
	.010	3.22	3.13	2.85	2.70	2.54	2.36	2.27	2.18
	.001	4.71	4.56	4.06	3.79	3.52	3.22	3.06	2.91
50	.100	1.76	1.73	1.63	1.57	1.50	1.42	1.38	1.33
	.050	2.07	2.03	1.87	1.78	1.69	1.58	1.51	1.45
	.025	2.38	2.32	2.11	1.99	1.87	1.72	1.64	1.56
	.010	2.78	2.70	2.42	2.27	2.10	1.91	1.80	1.70
	.001	3.82	3.67	3.20	2.95	2.68	2.38	2.21	2.05
100	.100	1.69	1.66	1.56	1.49	1.42	1.34	1.28	1.22
	.050	1.97	1.93	1.77	1.68	1.57	1.45	1.38	1.30
	.025	2.24	2.18	1.97	1.85	1.71	1.56	1.46	1.36
	.010	2.59	2.50	2.22	2.07	1.89	1.69	1.57	1.45
	.001	3.44	3.30	2.84	2.59	2.32	2.01	1.83	1.64
200	.100	1.66	1.63	1.52	1.46	1.38	1.29	1.23	1.16
	.050	1.93	1.88	1.72	1.62	1.52	1.39	1.30	1.21
	.025	2.18	2.11	1.90	1.78	1.64	1.47	1.37	1.25
	.010	2.50	2.41	2.13	1.97	1.79	1.58	1.45	1.30
	.001	3.26	3.12	2.67	2.42	2.15	1.83	1.64	1.43
1000	.100	1.64	1.61	1.49	1.43	1.35	1.25	1.18	1.08
	.050	1.89	1.84	1.68	1.58	1.47	1.33	1.24	1.11
	.025	2.13	2.06	1.85	1.72	1.58	1.41	1.29	1.13
	.010	2.43	2.34	2.06	1.90	1.72	1.50	1.35	1.16
	.001	3.13	2.99	2.54	2.30	2.02	1.69	1.49	1.22

TABLE G: **Critical values m* for Tukey pairwise multiple comparisons with 95% confidence level, *k* comparisons, and *N* – *k* degrees of freedom (df)**

df	Number of comparisons (*k*)							
	3	4	5	6	7	8	9	10
5	3.254	3.690	4.011	4.266	4.476	4.654	4.809	4.946
6	3.068	3.462	3.751	3.980	4.168	4.329	4.468	4.591
7	2.945	3.310	3.578	3.789	3.964	4.112	4.241	4.354
8	2.857	3.202	3.455	3.654	3.818	3.957	4.078	4.185
9	2.792	3.122	3.362	3.553	3.708	3.841	3.956	4.057
10	2.741	3.060	3.291	3.473	3.623	3.750	3.861	3.958
11	2.701	3.009	3.234	3.410	3.555	3.678	3.785	3.879
12	2.668	2.969	3.188	3.359	3.500	3.620	3.723	3.815
13	2.640	2.935	3.149	3.316	3.454	3.570	3.671	3.760
14	2.617	2.907	3.116	3.280	3.415	3.528	3.627	3.714
15	2.597	2.882	3.088	3.249	3.381	3.493	3.590	3.676
16	2.580	2.861	3.064	3.222	3.352	3.462	3.557	3.642
17	2.565	2.843	3.043	3.199	3.327	3.435	3.529	3.612
18	2.552	2.826	3.024	3.178	3.304	3.411	3.504	3.586
19	2.541	2.812	3.007	3.159	3.285	3.390	3.482	3.562
20	2.530	2.799	2.992	3.143	3.267	3.371	3.461	3.541
21	2.521	2.787	2.979	3.128	3.251	3.354	3.444	3.522
22	2.512	2.777	2.967	3.115	3.236	3.339	3.427	3.505
23	2.505	2.768	2.956	3.103	3.223	3.325	3.412	3.490
24	2.498	2.758	2.946	3.092	3.211	3.312	3.399	3.475
25	2.491	2.751	2.937	3.082	3.200	3.300	3.386	3.463
26	2.485	2.744	2.928	3.072	3.190	3.289	3.375	3.451
27	2.479	2.737	2.920	3.064	3.181	3.280	3.364	3.439
28	2.474	2.730	2.913	3.056	3.172	3.270	3.355	3.429
29	2.470	2.724	2.907	3.048	3.164	3.262	3.346	3.420
30	2.465	2.719	2.901	3.041	3.157	3.253	3.338	3.411
40	2.434	2.681	2.856	2.992	3.103	3.197	3.277	3.348
50	2.415	2.657	2.830	2.963	3.072	3.163	3.241	3.310
60	2.403	2.642	2.812	2.944	3.050	3.140	3.217	3.285
80	2.388	2.624	2.791	2.920	3.024	3.113	3.188	3.255
100	2.379	2.613	2.778	2.906	3.009	3.096	3.171	3.236
120	2.373	2.606	2.770	2.896	2.999	3.085	3.159	3.224
150	2.367	2.598	2.761	2.887	2.989	3.075	3.147	3.212
200	2.361	2.591	2.753	2.877	2.978	3.063	3.136	3.200

Answers to Selected Exercises

Chapter 1

1.1 **(a)** Cereal brands. **(b)** Categorical: manufacturer, preparation. Quantitative: calories, sugar content, fiber content.

1.3 **(a)** Individuals: children. Variable: preference. **(c)** Cartoon characters appear to influence cracker preference.

1.7 Single-peaked and right-skewed.

1.9 They are very similar. Because the dotplot shows the exact location of each data point, it makes it easier to spot and examine outliers.

1.11 The number of deaths has increased fairly steadily over time.

1.15 (a)

1.16 (c)

1.17 (c)

1.18 (a)

1.19 (b)

1.20 (b)

1.21 (c)

1.22 (a)

1.23 (b)

1.24 (b)

1.25 **(a)** Quantitative. **(b)** Quantitative. **(c)** Categorical. **(d)** Quantitative. **(e)** Categorical.

1.27 **(a)** Lakes. **(b)** 5 variables: 4 quantitative and 1 categorical (age of data).

1.29 **(a)** 130. **(b)** 20.7%. **(d)** Yes; together, these are all the categories of manatee deaths.

1.31 **(a)** All age groups have a fairly large rate of obesity. **(b)** No; each percent represents a different age group separately.

1.33 **(a)** Bar graph: categorical data. **(b)** Histogram: quantitative data. **(c)**

Histogram: quantitative data. **(d)** Bar graph: categorical data.

1.35 **(a)** The two separate peaks could reflect a bimodal distribution. **(b)** The second histogram appears unimodal, but with a wide peak.

1.37 **(a)** The distribution is roughly symmetric with no outliers. **(c)** Both graphs show all data points. The dotplot gives a more precise spatial representation while the stemplot provides the exact numerical values for easy reading.

1.39 **(b)** The distribution is skewed to the right.

1.41 **(a)** The distribution is skewed with nanoparticles, and somewhat symmetric with the buffer treatment. **(b)** The midpoints are approximately 6 with buffer, and 2 with nanoparticles. Tumor increase tends to be smaller with nanoparticle treatment.

1.43 **(a)** Increasing trend starting in the mid 1970s, but no cyclical fluctuations. **(b)** No; a histogram would mask the existence of the clear trend over time.

Chapter 2

2.1 **(a)** The distribution is irregular but somewhat symmetric except for a high outlier. The spread is from 164.0 to 740.2. **(b)** $\bar{x}$=319.25; $\bar{x} = 319.25;$ median=290.70.$\text{median} = 290.70.$ The median is notably less than the mean, which suggests a right-skew or high outlier(s).

2.3 **(a)** Median=6.$\text{Median} = 6.$ **(b)** $\bar{x}$=5.43.$\bar{x} = 5.43.$ **(c)** The age distribution is moderately left-skewed.

2.5 **(a)** Min.=164.0,$\text{Min.} = 164.0,$ Q1=260.9,$Q_1 = 260.9,$ M=290.7,$M = 290.7,$ Q3=354.95,$Q_3 = 354.95,$ max.=740.2.$\text{max.} = 740.2.$ **(b)** $\bar{x}$=319.25; $\bar{x} = 319.25;$ s=124.92.$s = 124.92.$ **(c)** The five-number summary is more informative. In particular, it reflects the fact that the maximum (an outlier) is much farther away from the median than the minimum.

2.7 **(a)** Both distributions are extremely skewed to the right, with fairly similar centers and spreads. **(b)** The mean and standard deviation are not resistant to skews and outliers.

2.9 740.2 is the only outlier.

2.11 **(a)** Factors 1 and 2 (incorrect use) and factors 3 and 4 (experimental consistency). **(b)** Factors 5 through 10 (natural variation in healthy humans). **(c)** Factors 1 through 4, which represent errors in experimentation or data collection. Additional factors may be argued, depending on purpose.

2.13 The two distributions overlap completely but scores tend to be higher, on average, for individuals with HSAM ($\bar{x}=8.1$)$(\bar{x} = 8.1)$ than for controls ($\bar{x}=5.2$).$(\bar{x} = 5.2)$.

2.14 (c)

2.15 (b)

2.16 (c)

2.17 (c)

2.18 (b)

2.19 (c)

2.20 (a)

2.21 (b)

2.22 (b)

2.23 (a)

2.25 The data are bimodal. There is no unique center and, therefore, no simple numerical summary.

2.27 $M=14$,$M = 14$, $Q1=13$, and $Q3=15$.$Q_1 = 13$, and $Q_3 = 15$.

2.29 **(a)** Symmetric distributions with no outliers. **(b)** On average, women have lower metabolic rates than men.

2.31 **(a)** Buffer: $\bar{x}=6.08$,$\bar{x} = 6.08$, $s=1.98$.$s = 1.98$. Nanoparticles: $\bar{x}=2.020$,$\bar{x} = 2.020$, $s=1.02$.$s = 1.02$. **(b)** The nanoparticle treatment was much more effective at limiting tumor growth.

2.33 **(a)** Any set of four identical numbers. **(b)** 0, 0, 10, 10 is the only possible answer.

2.35 Both distributions overlap and have very similar measures of center (whether assessed by their mean or by their median), so there is little reason to believe that the addition of cicadas affects seed mass, on average.

2.37 Here, $\bar{x}$ and $s$$\bar{x}$ and s are reasonable summaries. Group 1: $\bar{x}=23.75$,$\bar{x} = 23.75$, $s=5.07$.$s = 5.07$. Group 2: $\bar{x}=14.08$,$\bar{x} = 14.08$, $s=4.98$.$s = 4.98$. Group 3: $\bar{x}=15.78$,$\bar{x} = 15.78$, $s=5.76$.$s = 5.76$. Logging appears to reduce the number of trees per plot.

2.39 Aggressions received are both higher and more variable for females of lower ranks. The fourth dominance group has an extreme outlier.

2.41 **(a)** Min.=17,$\text{Min.} = 17,$ Q1=22.5,$Q_1 = 22.5,$ M=23.5,$M = 23.5,$ Q3=25, $Q_3 = 25,$ max.=31.5.$\text{max.} = 31.5.$ **(b)** IQR=2.5,$IQR = 2.5,$ 1.5×IQR=3.75; $1.5 \times IQR = 3.75;$ observations less than 18.75 or greater than 28.75 are suspected outliers. Only DC and Utah are flagged. **(c)** x̄=23.755, $\bar{x} = 23.755,$ very close to M=23.5$M = 23.5$ (distribution is symmetric).

Chapter 3

3.1 **(a)** Explanatory: number of calories. Response: percent body fat. **(b)** Explore the relationship. **(c)** Explanatory: inches of rain. Response: yield of corn. **(d)** Explore the relationship.

3.3 **(a)** Neither is clearly the explanatory variable, and each one may influence the other to some extent. **(b)** The stated objective is to study opinions as the response variable, making emissions the explanatory variable in this context.

3.5 **(b)** Linear, negative, and reasonably strong. The number of D notes per call appears to represent predator wingspan.

3.7 **(a)** Mass is explanatory. Subject number is an index variable and thus contains no data. **(b)** Moderately strong, positive linear relationship. **(c)** The same relationship seems to hold. Men tend to have higher values and are overall more variable than women.

3.9 The relationship is clearly not linear; therefore, *r* is meaningless.

3.11 (a)

3.12 (a)

3.13 (a)

3.14 (b)

3.15 (c)

3.16 (c)

3.17 (a)

3.18 (a)

3.19 (b)

3.20 (c)

3.21 **(a)** The negative linear relationship suggests that it is. **(b)** r=−0.928. $r = -0.928$. This is very strong.

3.23 **(a)** The lack of association suggests that it isn't. **(b)** r=0.115. $r = 0.115$. This is very weak. **(c)** *D. magna* may be an effective method but not *D. pulex*.

3.25 **(b)** Moderate, negative linear association (r=−0.5436).$(r = -0.5436)$. **(c)** No, this is a complex problem with no obvious explanatory variable and likely multiple influential factors.

3.27 **(a)** Linear, positive, and very strong (r=0.9766).$(r = 0.9766)$. **(b)** The 3D reconstruction is a better model (*r* stronger).

3.29 Strong, positive linear relationship (r=0.8117).$(r = 0.8117)$.

3.31 **(a)** The negative linear association indicates habituation. **(b)** The patterns are different for the mutant and the wild type, suggesting that PVD sensory neurons are involved in nociception and habituation.

3.33 **(a)** Most of the data are lumped in the lower right corner, so *r* could be misleading. **(b)** The data range from 0.00003 to 1.0 for lifetime risk and from 3,150,000 to 3,550,000,000,000 for stem cell divisions. In log scale, however, the ranges are −4.52-4.52 to 0 and 6.5 to 12.55, respectively. **(c)** The scatterplot now shows a strong, positive linear relationship (r=0.804,$r = 0.804$, up to roundoff error).

3.35 Moderately strong, positive linear association (r=0.8042).$(r = 0.8042)$. This supports the idea that animal populations decline when the fish supply is low.

3.37 Strong, positive linear association (r=0.9143).$(r = 0.9143)$. This supports the idea that beak size may play a role in cooling.

3.39 **(a)** r=1$r = 1$ for a line. **(c)** Leave some space above your vertical stack. **(d)** The curve must be higher at the right than at the left.

Chapter 4

4.1 **(a)** The slope indicates that for each drink consumed, BAC increases by 0.023, on average. The intercept is zero because someone who doesn't have any drinks should have a BAC of 0. **(b)** BAC=0.038× number of drinks.$\text{BAC} = 0.038 \times \text{number of drinks.}$ **(c)** BAC=0.023×4=0.092,$\text{BAC} = 0.023 \times 4 = 0.092,$ on average, for a man consuming 4 drinks; BAC=0.038×2=0.076,$\text{BAC} = 0.038 \times 2 = 0.076,$ on average, for a woman consuming 2 drinks.

4.3 **(a)** $\bar{x}$=30.261,$\bar{x} = 30.261,$ sx=0.4391,$s_x = 0.4391,$ $\bar{y}$=2.511,$\bar{y} = 2.511,$ sy=0.1502,$s_y = 0.1502,$ r=−0.8986,$r = -0.8986,$ b=rsy⁄sx=−0.3074 and a=$\bar{y}$-b$\bar{x}$=11.813.$b = rs_y/s_x = -0.3074$ and $a = \bar{y} - b\bar{x} = 11.813.$ **(b)** $\hat{y}$=11.811−0.3073x.$\hat{y} = 11.811 - 0.3073x.$

4.5 **(a)** Starvation rate explained 70% of the variation in fledging rate. **(b)** r=−0.84.$r = -0.84.$

4.7 **(a)** Using all points, r=−0.684.$r = -0.684.$ Without the outliers, r= −0.836.$r = -0.836.$ **(b)** Using all points, y=61.39−0.6806x.$y = 61.39 - 0.6806x.$ Without the outliers, y=68.18−0.7883x.$y = 68.18 - 0.7883x.$ **(c)** They are outliers of the relationship and are influential especially for the correlation.

4.9 **(a)** LogRisk=−7.6108+0.5326$\text{LogRisk} = -7.6108 + 0.5326$LogDivisions. **(b)** Predicted LogRisk=−2.8174.$\text{Predicted LogRisk} = -2.8174.$ **(c)** Predicted risk=0.00152.$\text{Predicted risk} = 0.00152.$

4.11 **(a)** Time since ingestion (explanatory) and zolpidem blood level (response). **(b)** Sex (women eliminate the drug more slowly).

4.13 **(a)** No; there is no biological reason why education would cause breast cancer. **(b)** Women with higher education tend to have babies later in life.

4.15 (b)

4.16 (c)

4.17 (c)

4.18 (a)

4.19 (a)

4.20 (a)

4.21 (a)

4.22 (b)

4.23 (c)

4.24 (c)

4.25 **(a)** Slope=101.98,$\text{Slope} = 101.98,$ positive relationship. **(b)** Prediction 49.8,$49.8,$ very reliable with r2=0.95.$r^2 = 0.95.$ **(c)** Nonsense prediction 131.4;$131.4;$ extrapolation (prediction outside range).

4.27 **(a)** Both relationships are linear but the relationship is stronger with the 3D approach. **(b)** y=0.3367+0.2649x$y = 0.3367 + 0.2649x$ so, for x=35,$x = 35,$ ŷ=9.608 cm3.$\hat{y} = 9.608\,\text{cm}^3.$ **(c)** y=0.419572+2.475224x $y = 0.419572 + 2.475224x$ so, for x=4.5,$x = 4.5,$ ŷ=11.558 cm3. $\hat{y} = 11.558\,\text{cm}^3.$ **(d)** r2=0.846$r^2 = 0.846$ for 2D; r2=0.954$r^2 = 0.954$ for 3D. The 3D approach is a more accurate predictive model. **(e)** The regression equation alone is not sufficient to judge a model.

4.29 **(a)** Slope=0.547.$\text{Slope} = 0.547.$ Intercept=2.791.$\text{Intercept} = 2.791.$ **(b)** Prediction=30.2.$\text{Prediction} = 30.2.$ A high r2r^2 (here, 0.9216) implies a reliable prediction.

4.31 **(a)** y=0.06078x−0.1261.$y = 0.06078x - 0.1261.$ Prediction=−0.00454. $\text{Prediction} = -0.00454.$ **(b)** 77.1%.

4.33 **(a)** They all give similar results; y=3.0+0.5x;$y = 3.0 + 0.5x;$ r=0.82; $r = 0.82;$ very close slopes, intercepts, and correlations; ŷ=8$\hat{y} = 8$ for x=10.$x = 10.$ **(c)** Only the first relationship is linear and appropriate for regression. The second relationship is strongly curved. The third has a strong outlier. The last has only two values for x.$x.$

4.35 **(a)** Ignoring the outlier, the plot shows a moderately strong, negative linear relationship. **(b)** For all 12 points: r=−0.3387. $r = -0.3387.$ Without the outlier: r=−0.7866.$r = -0.7866.$ The outlier weakens the linear pattern. **(c)** For all 12 points: ŷ=−492.6−33.79x. $\hat{y} = -492.6 - 33.79x.$ Without the outlier: ŷ=−1371.6−75.52x. $\hat{y} = -1371.6 - 75.52x.$ The line is pulled toward the outlier, reducing the very large vertical deviation of this point from the line.

4.37 **(a)** ŷ=2.537.$\hat{y} = 2.537.$ **(b)** Predicted mean seed count=102.537=344.3.$\text{Predicted mean seed count} = 10^{2.537} = 344.3.$

4.39 No; it is more probable that those in critical condition are sent to large hospitals.

4.41 **(a)** No, association does not imply causation; there are likely confounding variables. **(b)** Answers will vary. They might have a lower socioeconomic status that limits their access to healthy food and health care, for instance.

4.43 The association is strong: Strength of association. The association is consistent: Consistency. Higher doses are associated with stronger responses: Biological gradient. The alleged cause precedes the effect in time: Temporality. The alleged cause is plausible: Biological plausibility and coherence, Biological mechanisms of HPV carcinogenesis. The five criteria are met.

4.45 Only the first scatterplot of 10 points from part (b).

Chapter 5

5.1 **(a)** Peanut exposure and peanut allergy. **(b)** 98 infants were enrolled, 51 of whom avoided peanuts. **(c)** Still allergic: 23 (23.5%). No longer allergic: 75 (76.5%).

5.3 There are an infinite number of possibilities, but a=50,$a = 50,$ b=0, $b = 0,$ c=10,$c = 10,$ d=40, and a=40,$d = 40, \text{ and } a = 40,$ b=10,$b = 10,$ c=20, and d=30$c = 20, \text{ and } d = 30$ are two solutions.

5.5 **(a)** Conditional distribution of alarm call: given predator nearby = 26.6%; given predator far = 41.4%. **(b)** When a predator is close, prairie dogs raise the alarm 26.6% of the time. When a predator is far away (making it safer for the individual to give a warning call), the prairie dogs raise the alarm 41.4% of the time.

5.7 **(a)** Open surgery: 78.0%. PCNL: 82.6%. **(b)** Small stones: 88.2%. Large stones: 72.0%. **(c)** Open surgery success for small stones: 93.1%. PCNL success for small stones: 86.7%. Open surgery success for large stones: 73.0%. PCNL success for large stones: 68.8%. This is Simpson's paradox: PCNL was used more often to treat small stones, which appear to be easier to treat (better chance of successful outcome).

5.9 (b)

5.10 (c)

5.11 (a)

5.12 (c)

5.13 (b)

5.14 (a)

5.15 (c)

5.16 (a)

5.17 (b)

5.18 (b)

5.19 **(a)** Marginal for depression: treated=11%,$\text{treated} = 11\%$, not treated=89%.$\text{treated} = 89\%$. **(b)** These make the conditional distribution of being treated, given different age groups, so they don't have to add to 100%.

5.21 **(a)** Overall, 60.4% success ("Yes") and 39.6% failure ("No"). **(b)** Conditional distributions: percent successful outcome was 80.8% in the Botox group but only 40.7% in the placebo group. Botox was much more effective.

5.23 **(a)** Percent in 9- to 11-month age range who ate fried potatoes: 9.0%. Percent in 15- to 18-month range: 20.1%. Percent in 19- to 24-month age range: 25.9%. **(b)** Association does not imply causation.

5.25 Conditional distribution of birth outcomes, given vitamins and minerals: 1.3% with congenital malformation, 0.0% with neural-tube defect, and 98.7% with neither. Conditional distribution of birth outcomes, given minerals only: 2.3% with congenital malformation, 0.3% with neural tube defect, and 97.4% with neither. Outcomes are more favorable with vitamins and minerals.

5.27 **(a)** Conditional distribution of successful outcomes, given treatment: 0.6% with placebo, 21.7% with ustekinumab, 25.7% with brodalumab low dose, and 44.4% with brodalumab high dose. **(b)** All three treatments performed better than the placebo, indicating an effect of the drug itself. **(c)** Brodalumab high dose was the most effective treatment.

5.29 **(a)** 99.2% for type 1 and 100% for type 2 among younger patients; 54.4% for type 1 and 58.8% for type 2 among older patients. Overall, younger patients have a much higher survival rate. Within each age group, patients with type 2 diabetes have a slightly higher survival rate. **(b)** 70.7% for type 1; 59.9% for type 2. It seems that patients with type 1 diabetes have a higher survival rate overall. **(c)** The lurking variable here is age. Type 2 diabetes is more frequent in older patients. This is Simpson's paradox.

5.31 **(b)** Percents hatching in each cell group: cold, 59.3%; neutral, 67.9%; hot, 72.1%. The cold temperature made hatching less likely.

5.33 Conditional distribution for dead trees: 55.9% western hemlock, 38.4% Douglas fir, and only 5.7% western red cedar. Conditional distribution for live trees: 43.8% western hemlock, 33.8% Douglas fir, and 22.3% western red cedar. Conditional distribution for sapling trees: 63.1% western red cedar, 36.1% western hemlock, and only 0.8% Douglas fir. The western red cedar is in the process of displacing the Douglas fir.

5.35 Conditional distribution of no smoking given treatment: 44.0% (Chantix), 29.5% (bupropion), 17.7% (placebo). Chantix has the best success rate after 9 weeks; the placebo group has the lowest success rate.

Chapter 6

6.1 Observational. The study did not attempt to change the participants in any way.

6.3 **(a)** Observational. **(b)** BMI, obesity, hydration status. **(c)** Water content of food consumed. **(d)** No, this is just an observational study.

6.5 Population.

6.7 **(a)** Label each plot from 0001 to 1410: plots labeled 769, 1315, and 94 are selected.

6.9 A multistage design divides the population into groups, takes a random sample of these groups, and then takes a random sample of individuals within the groups sampled. An SRS just samples from the entire population, without dividing it into subgroups. Both techniques use randomness for selecting individuals.

6.11 **(a)** The population is all physicians practicing in the United States. The sample size is 2379 (physicians for whom we do have data). Conclusions apply to all U.S. physicians. **(b)** The rate of nonresponse is 97.6%. There is a very real potential for bias. **(c)** The Physicians Foundation hoped to survey 100,000 physicians but obtained data from only 2379 physicians. The report should be called "a survey of 2379 physicians."

6.13 This is a cohort study. The population is full-term babies born in a similar time and place. The explanatory variable is antibiotic exposure. The response variable is body mass recorded over time.

6.15 (c)

6.16 (c)

6.17 (c)

6.18 (a)

6.19 (b)

6.20 (a)

6.21 (b)

6.22 (b)

6.23 (a)

6.24 (c)

6.25 **(a)** A convenience sample. The population is Jawbone UP users. **(b)** Nothing beyond the stated summary statistics representing this particular convenience sample.

6.27 In a stratified random sample, the entire population is broken up into subgroups; then an SRS is taken from each of the individual subgroups; and the SRSs are finally recombined to create the sample. This ensures that all four forest types are represented in the sample in proportion to their representation in the population.

6.29 **(a)** 56% and 43.8%, respectively. **(b)** The polls represent two different geographic areas with a different mix of individuals. The question wording is very different in the two polls.

6.31 **(a)** To avoid response bias due to the disturbance of a new presence. **(b)** If animals are scared or offended by a new, unusual presence, then we cannot observe their natural behavior.

6.33 **(a)** The question is clear, but the wording seems to suggest that cell phone usage caused the brain cancer. **(b)** The question is slanted toward an affirmative answer. **(c)** This question is so unclear and charged with negative words that it is likely to prevent many "yes" answers.

6.35 Answers will vary.

6.37 **(a)** All adults living in the United States. Some people have only a landline phone; others have only a cell phone. Choosing based on birthday is random and unbiased. Selecting the one who answers might introduce a bias if some types of people tend to answer the phone more often. **(b)** If different types of people are easier to reach, they would be over-represented in a random sample. Weights can compensate for such disparities but not for answer bias (inaccurate or untruthful answers).

6.39 **(a)** A cohort study. One homogeneous group was followed for many years to track the effect of caffeine on depression. **(b)** This is an observational study, so a causal link cannot be established. **(c)** Answers will vary. Nurses with health problems could avoid caffeine, and health problems could also lead to depression and antidepressant use.

6.41 **(a)** A cohort study. **(b)** There are enough cases for statistical evaluation. **(c)** No, this is an observational study.

6.43 **(a)** A cohort study. **(b)** No, this is an observational study. **(c)** For example, existing health concerns may force the person to quit smoking.

Chapter 7

7.1 Subjects: the healthy people aged 18 to 40. Factor: the pill given to the subject. Treatments: ginkgo or placebo. Response variable: the number of *e*’s missed by each subject.

7.3 **(a)** Randomization prevents bias in the assignment to conditions. **(b)** Without a control group serving as comparison, whatever effect would be found with POM consumption could be completely unrelated to POM consumption. **(c)** If a positive outcome occurs whether men consumed POM or a placebo, then the outcome cannot be attributed to POM consumption.

7.5 **(a)** Assign 6 plots to each treatment. **(b)** If using Table A, label 01 to 36 and take two digits at a time.

7.7 **(a)** The individuals are infants with severe eczema. The treatments are peanut consumption and peanut avoidance. The response variable is peanut allergy at 60 months of age. **(b)** Block design. **(c)** A pre-existing peanut allergy could be influential, so to have these subgroups evenly represented in both treatments.

7.9 **(a)** Randomly assign 15 students to Group 1 (easy mazes) and the other 15 to Group 2 (hard mazes). Compare each group's time estimates. **(b)** Each student does the activity twice, once with the easy mazes and once with the hard mazes, in a randomly determined order. Compare each student's easy and hard time estimates **(c)** Randomly assign half the men and half the women to Group 1, and the remaining subjects to Group 2. Compare each group's time estimates.

7.11 **(a)** The individuals are patients with PAD. The treatments are milk chocolate and dark chocolate. The response variable is maximum walking distance. **(b)** Matched pairs design (repeated measures). Single blind (subjects can taste the difference between dark chocolate and milk chocolate).

7.13 Answers will vary. **(a)** One difference is the inability to refuse participation. **(b)** Not a truly informed consent. **(c)** No. The experimenters did not know when they started the experiment that the effects would be small. **(c)** In this case, the scientific community voiced its concern clearly and immediately.

7.14 (b)

7.15 (b)

7.16 (c)

7.17 (a)

7.18 (a)

7.19 (c)

7.20 (b)

7.21 (c)

7.22 (b)

7.23 (a)

7.25 **(a)** Explanatory: type of operation. Response: survival time. **(b)** Not an experiment: existing records are simply examined. **(c)** The type of operation is most likely decided based on the size of the tumor. Larger tumors would be much more dangerous to remove, regardless of type of operation.

7.27 The first study is an observational, case-control study; the second is an experiment (the exercise plan is assigned randomly).

7.29 **(a)** Completely randomized design: 12 birds randomly selected for the placebo group, with the remaining 15 making up the immune challenge group (or the other way around). **(b)** Using Table A to select the 12 birds in the placebo group: 05, 19, 04, 25, 20, 16, 18, 07, 13, 02, 23, 27.

7.31 **(a)** Randomly assign the subjects to four groups: Group 1, antidepressants and stress management; Group 2, stress management only; Group 3, antidepressants only; Group 4, control. **(b)** Using Table A, Group 1 = Chai, Hammond, Herrera, Xiang, Irwin, Hurwitz, Reed, Broden, Lucero, Nho.

7.33 **(a)** Each subject was scanned twice (ON and OFF conditions). **(b)** Different subjects might have different brain activities. The matched pairs design allows comparison of the ON and OFF conditions for each subject. **(c)** The difference could not easily be explained by chance variations alone.

7.35 **(a)** A block design.

7.37 **(a)** Randomly sample half of the subjects and give them the treatment; everyone else gets the placebo pill. Then compare the results. **(b)** Using Table A: 170, 005, 227, 118, 007.

7.39 **(a)** This is a nonrandomized, noncomparative study. **(b)** No, because subjects were not randomly selected (they were hand-picked by Lipov) and the lack of comparison means that results could be a placebo effect.

7.41 **(a)** In an observational study, we observe subjects who have chosen to take supplements and compare them with others who do not take supplements. In an experiment, we assign some subjects to take supplements and others to take a placebo. **(b)** Treatments are assigned at random, and a control group is used as a basis for comparison to observe the effects of the treatment. **(c)** Subjects who choose to take supplements are more likely to make healthy lifestyle choices. When random assignment is used, some of those subjects will take the supplement and some will take the placebo.

7.43 “No significant difference” does not mean the groups are identical but rather that observed differences are no bigger than we might expect from true random allocation alone.

7.45 **(b)** Yes; the results appear trustworthy. The study was a comparative experiment with a placebo control group, subjects were randomly assigned to treatments, double-blinding was used to prevent bias, and the variable was clearly defined.

7.47 **(a)** The procedure might be harmful in humans. **(b)** It might be ethical to test the procedure without consent if it was believed that the patient would not survive otherwise.

7.49 The Tanzanian government and the researchers both presented serious arguments. However, the study’s purpose was entirely academic. Without patients’ informed consent or any benefit to them, the study was not ethically justified.

7.51 For example, informed consent is lacking.

7.53 Exact outcomes will vary.

Chapter 8

8.1 **(a)** Mildly left-skewed. **(b)** 3,000 to 3,499 grams. **(c)** Very-low 1.4%, low 6.7%, normal 91.9%.

8.3 **(b)** The distribution is clearly bimodal, so a single summary (mean and standard deviation or the five number summary) would be misleading. **(c)** Left-biased 42%, ambidextrous 19%, right-biased 39%.

8.5 **(a)** The distribution is irregular but roughly symmetric: $\bar{x}$=15.0, $\bar{x} = 15.0,$ s=6.29.$s = 6.29.$ **(b)** Lots of variation but with an overall downward trend. **(c)** Day=164.7−0.07610$\text{Day} = 164.7 - 0.07610$Year; r2=0.126.$r^2 = 0.126.$

8.7 **(a)** Roughly symmetric with a high outlier (HAV=50).$(\text{HAV} = 50).$ With the entire set: median=25,$\text{median} = 25,$ $\bar{x}$=25.42,$\bar{x} = 25.42,$ s=7.47.$s = 7.47.$ Without the outlier: median=25,$\text{median} = 25,$ $\bar{x}$=24.76,$\bar{x} = 24.76,$ s=6.34.$s = 6.34.$ **(b)** Right-skewed. Median=16,$\text{Median} = 16,$ $\bar{x}$=16.82,

$\bar{x} = 16.82,$ s=6.67.$s = 6.67.$ **(c)** y=0.339x+19.723.$y = 0.339x + 19.723.$ **(d)** Not very well; the association is fairly weak and r2=0.0913. $r^2 = 0.0913.$

8.9 **(a)** Right-skewed with a high outlier; median=72,$\text{median} = 72,$ x̄=106.2,$\bar{x} = 106.2,$ s=91.8,$s = 91.8,$ (x̄=96.0($\bar{x} = 96.0$ without the outlier). **(b)** Right-skewed; median=141,$\text{median} = 141,$ x̄=176.6,$\bar{x} = 176.6,$ s=111.8.$s = 111.8.$ **(c)** 31 neurons, 84%. **(d)** r=0.6386;$r = 0.6386;$ the relationship is moderately strong.

8.11 **(a)** Both distributions are strongly right-skewed with extreme outliers. **(b)** Both transformed distributions are roughly symmetric. **(c)** The scatterplot for raw weights is mostly in the lower left corner. The scatterplot for transformed weights shows a clear positive, linear association. **(d)** r2=0.90.$r^2 = 0.90.$

Chapter 9

9.1 This means that if you repeatedly sampled 100,000 males, on average you would find 13 who were diagnosed with hemophilia. For one random sample of 100,000 males, you could find more or you could find fewer than 13 individuals with hemophilia.

9.3 **(a)** 0. **(b)** 1. **(c)** 0.01. **(d)** 0.6.

9.5 **(a)** S={BBB, BBG, BGB, BGG, GBB, GBG, GGB, GGG}. $S = \{\text{BBB}, \text{BBG}, \text{BGB}, \text{BGG}, \text{GBB}, \text{GBG}, \text{GGB}, \text{GGG}\}.$ **(b)** S={0,1,2,3}. $S = \{0, 1, 2, 3\}.$ Associated probabilities=18,38,38,and18, $\text{Associated probabilities} = \frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \text{and} \frac{1}{8},$ respectively.

9.7 **(a)** 0.15. **(b)** 0.47. **(c)** 0.32.

9.9 **(a)** 0.48. **(b)** 0.89. **(c)** 1.00.

9.11 **(a)** 0.4. **(b)** 0.4. **(c)** 0.2. **(d)** 0.8.

9.13 **(a)** The sum of probabilities is 1. **(b)** The student was physically active on fewer than 7 days in the past week. P(X<7)=0.73. $P(X < 7) = 0.73.$ **(c)** P(X>0)=P(X≥1)=0.85.$P(X > 0) = P(X \geq 1) = 0.85.$ **(d)** μ=3.92,$\mu = 3.92,$ σ=2.54.$\sigma = 2.54.$

9.15 **(a)** Probability=0.268,$\text{Probability} = 0.268,$ risk=0.268,$\text{risk} = 0.268,$ odds=0.366.$\text{odds} = 0.366.$ **(b)** By age group of increasing age, risk: 0.245, 0.448, 0.274, 0.275, 0.252, and 0.196, respectively; odds:

0.325, 0.812, 0.377, 0.379, 0.337, and 0.244. Risk and odds are greatest among 20- to 24-year-olds.

9.16 (a)

9.17 (b)

9.18 (b)

9.19 (c)

9.20 (b)

9.21 (c)

9.22 (b)

9.23 (b)

9.24 (b)

9.25 (c)

9.27 **(a)** 0.2066. **(b)** 0.7934. **(c)** 0.3801 and 0.6199, respectively.

9.29 **(a)** S={accident, homicide, suicide, other}. $S = \{\text{accident, homicide, suicide, other}\}$. **(b)** 0.72 and 0.28, respectively.

9.31 **(a)** 0.04. **(b)** 0.75.

9.33 **(a)** The sum of probabilities is 1. Discrete. **(b)** 0.287. **(c)** 0.713.

9.35 **(a)** 0.14. **(b)** 0.43. **(c)** 0.75.

9.37 The groups make up all possible age outcomes and the sum of probabilities is 100%. Discrete. **(b)** 53%.

9.39 **(a)** X<2$X < 2$ or X>4.$X > 4$. P(X<2$P(X < 2$ or X>4)=0.17.$X > 4) = 0.17$. **(b)** “Eye color greater than 2 and less than or equal to 4.” P(2<X≤4)=0.72.$P(2 < X \leq 4) = 0.72$.

9.41 **(a)** At most one site: 0.48. P(2≤X<4)=0.40.$P(2 \leq X < 4) = 0.40$. **(b)** μ=1.76,$\mu = 1.76$, the mean number of social media used (of the main 5) we’d expect to find among very many American adult internet users. **(c)** σ=1.36,$\sigma = 1.36$, which can be thought of, roughly speaking, as the expected deviation from the mean in this population.

9.43 **(a)** YY is continuous, because it can take any value between 0 and 2. **(b)** Height=0.5.$\text{Height} = 0.5$. **(c)** P(Y≤1)=0.5.$P(Y \leq 1) = 0.5$.

9.45 **(a)** P(13≤X<16)=0.77.$P(13 \leq X < 16) = 0.77$. **(b)** P(X>13)=0.88. $P(X > 13) = 0.88$.

9.47 **(a)** Risk=0.0043,$\text{Risk} = 0.0043,$ odds=0.00432.$\text{odds} = 0.00432.$ **(b)** Risk=0.0365,$\text{Risk} = 0.0365,$ odds=0.0379.$\text{odds} = 0.0379.$ **(c)** Probability=risk=0.127,$\text{Probability} = \text{risk} = 0.127,$ odds=0.145.$\text{odds} = 0.145.$ **(d)** By decade, because risk varies over a lifetime.

9.49 Actual results will vary.

Chapter 10

10.1 **(a)** P(next two patients are males)=0.5625. $P\,(\text{next two patients are males}) = 0.5625.$ **(b)** P(at least one is female)=0.7627.$P\,(\text{at least one is female}) = 0.7627.$ P(at least one is male)=0.9990.$P\,(\text{at least one is male}) = 0.9990.$

10.3 **(a)** 0.055. **(b)** 0.945.

10.5 $P(\text{CDK}) = 0.091$. $P(\text{CDK} \mid \text{diabetes}) = 0.141$. $P(\text{CDK} \mid \text{no diabetes}) = 0.064$.

10.7 **(b)** P(I)=0.287.$P\,(I) = 0.287.$ P(OM)=0.716.$P\,(OM) = 0.716.$ **(c)** 0.856. **(d)**P(I)×P(OM)=0.205≠0.147,$P\,(I) \times P\,(OM) = 0.205 \neq 0.147,$ so insulin injections and oral medication are not used independently in this population.

10.9 **(a)** $P(\text{over 65}) = 1/4$. $P(\text{actively practicing} \mid \text{over 65}) = 2/5$. **(b)** 1/10.

10.11 **(a)** $P(\text{blue eyes} \mid \text{red hair}) = 0.473$. $P(\text{blue eyes and red hair}) = 0.0118$. **(b)** $P(\text{freckles} \mid \text{red hair and blue eyes}) = 0.857$. $P(\text{red hair and blue eyes and freckles}) = 0.0101$.

10.13 **(a)** $P(\text{red hair}) = 0.025$, $P(\text{blue eyes}) = 0.388$, $P(\text{freckles}) = 0.253$. **(b)** $P(\text{freckles and red hair}) = 0.0203$. $P(\text{freckles} \mid \text{red hair}) = 0.811$.

10.15 **(a)** 24.5%. **(b)** 0.233.

10.17 0.233.

10.18 (a)

10.19 (c)

10.20 (c)

10.21 (c)

10.22 (b)

10.23 (b)

10.24 (a)

10.25 (b)

10.26 (c)

10.27 (b)

10.29 $P(X=0)=0.0282$.$P(X=0) = 0.0282$. $P(X\geq 1)=0.9718$.$P(X \geq 1) = 0.9718$.

10.31 Because having brown hair and having freckles are not necessarily independent events. (The earlier tree diagram shows that they are definitely not independent.)

10.33 84%.

10.35 **(a)** 0.2400. **(b)** 0.2844, 0.3248, 0.1991, and 0.1415 (top to bottom). **(c)** The probability of damage decreases noticeably when thorny cover is 1/3 or more.

10.37 **(a)** 0.0651. **(b)** 0.0605. **(c)** Fourth=0.0563,$\text{Fourth} = 0.0563,$ fifth=0.0524,$\text{fifth} = 0.0524,$ kth=$k\text{th} =$ (0.07)(0.93)k−1.$(0.07)(0.93)^{k-1}$.

10.39 $P(X=0)=0.9039$,$P(X=0) = 0.9039,$ $P(X=1)=0.0922$, and $P(X=2)=0.0038$;$P(X=1) = 0.0922, \text{ and } P(X=2) = 0.0038;$ the other three probabilities are less than 0.0001.

10.41 **(a)** *P*(both carriers | both of European Jewish descent) = 0.00137, *P*(both carriers | mother of European Jewish descent but father not) = 0.000148, *P*(both carriers | neither of European Jewish descent) = 0.000016. **(b)** *P*(child TS | both parents of European Jewish descent) = 0.00034. **(c)** *P*(child TS | neither parent of European Jewish descent) = 0.000004.

10.43 1.

10.45 0.194.

10.47 **(a)** 0.9292. **(b)** 0.9998. **(c)** 0.0256 and 0.9906, respectively. **(d)** PPV depends on the rate of HIV in the population screened because it influences how many true-positives and false-positives we would get. But the sensitivity and the specificity are properties of the OraQuick test that depend only on chemistry and manufacturing.

10.49 **(b)** *P*(yes) = *P*(yes and head) + *P*(yes and tail) = 0.5p + 0.5. **(c)** $p = (0.59/0.5) - 1 = 0.18$.

Chapter 11

11.1 **(a)** The men’s distribution is wider, so it must be shorter so that the total area is still equal to 1.

11.3 **(a)** 350 – 750 ml. **(b)** 16%. **(c)** 175 ml – 625 ml. **(d)** 250 ml and less.

11.5 At age 5, *z*-score = 0.556. At age 7, *z*-score = 0.476. John has been growing slightly more slowly than average, because his *z* score has gone down a little bit. A third measurement one or two years later would help our conclusion.

11.7 **(a)** 0.150. **(b)** 0.191. **(c)** 0.191.

11.9 **(a)** 678.2 ml and larger. **(b)** 523.4 ml and larger. **(c)** 482.6 ml for men and 349.4 ml for women.

11.11 **(a)** 0.0022. **(b)** 0.9493. **(c)** **(c)** z is between −0.67 and −0.68 (software gives $z = -0.674$). **(d)** z is between 0.25 and 0.26 (software gives $z = 0.253$).

11.13 It is close to Normal, because the dots appear to make a straight line.

11.15 (a)

11.16 (a)

11.17 (b)

11.18 (a)

11.19 (b)

11.20 (c)

11.21 (b)

11.22 (b)

11.23 (b)

11.24 (a)

11.25 **(a)** 50%, 0.15%, 16%, respectively. **(b)** 2.5%. **(c)** 2.5%.

11.27 **(a)** 1.22%. **(b)** 98.78%. **(c)** 3.84%. **(d)** 94.94%.

11.29 **(a)** $z=0.842$.$z = 0.842$. **(b)** $z=0.385$.$z = 0.385$.

11.31 Approximately $z>1$$z > 1$ for A, $0<z<1$$0 < z < 1$ for B, $-1<z<0$ $-1 < z < 0$ for C, $z<-1$$z < -1$ for D and F grades.

11.33 **(a)** 0.0267. **(b)** 0.00001.

11.35 0.73 – 5.67 μg/dl.

11.37 **(a)** 0.776. **(b)** 45.0%. **(c)** 4.18 bl/s and 8.22 bl/s.

11.39 **(a)** 59.87%. **(b)** 54.7%. **(c)** > 279 days.

11.41 **(a)** 97.42 to 99.78. **(b)** 96.83 to 99.57.

11.43 **(b)** 0.1%. **(c)** 14.1%. **(d)** The Normal approximations are quite accurate.

11.45 About 0.675, both above and below.

11.47 Atlantic acorn sizes are clearly not Normal (right-skew).

11.49 The data appear roughly Normal.

11.51 The plot for never-logged areas is nearly linear, indicating a roughly Normal distribution. The other two plots each show a low value, suggesting perhaps a slight skew to the left or a low outlier.

Chapter 12

12.1 Yes. Reach a live person or not. n=15, p=0.2.$n = 15,\ p = 0.2.$

12.3 Yes. n=100, p=0.44.$n = 100,\ p = 0.44.$

12.5 **(a)** Binomial. XX is resistance; n=10, p=0.27.$n = 10,\ p = 0.27.$ **(b)** P(X=1)=0.159,$P(X = 1) = 0.159,$ P(X=2)=0.265.$P(X = 2) = 0.265.$ **(c)** P(X≥1)=1-P(X=0)=0.957.$P(X \geq 1) = 1 - P(X = 0) = 0.957.$

12.7 **(a)** μ=2.7,$\mu = 2.7,$ σ=1.404.$\sigma = 1.404.$ **(b)** μ=3.8,$\mu = 3.8,$ σ=1.535. $\sigma = 1.535.$ **(c)** The standard deviation σσ is greater for the larger *p* (in 2014). For p=0.5,$p = 0.5,$ σ=1.581.$\sigma = 1.581.$

12.9 **(a)** Fixed n,$n,$ constant p,$p,$ and independent observations. **(b)** p=1⁄5, n=25, μ=5.$p = 1/5,\ n = 25,\ \mu = 5.$ **(c)** In a classroom of 25 students, *we expect* five of them, *on average*, to be struggling with a mental health disorder.

12.11 **(a)** P(X≤82)=0.0129.$P(X \leq 82) = 0.0129.$ **(b)** P(X≤82)=0.0145. $P(X \leq 82) = 0.0145.$ This is fairly close. **(c)** It is likely that many of the cases of osteopenia have been missed.

12.13 Poisson distribution (independent observations with a constant rate per year).

12.15 **(a)** {0,1,2,3,…}.$\{0, 1, 2, 3, \ldots\}.$ **(b)** Poisson with a rate of 5.7 cases per week. **(c)** μ=5.7, σ=2.387.$\mu = 5.7,\ \sigma = 2.387.$

12.19 (b)

12.20 (b)

12.21 (c)

12.22 (b)

12.23 (b)

12.24 (c)

12.25 (a)

12.26 (c)

12.27 (c)

12.28 (b)

12.29 **(a)** Binomial is reasonable, n=8.$n = 8.$ **(b)** Binomial is reasonable, n=100.$n = 100.$ **(c)** Binomial is reasonable, n=24.$n = 24.$

12.31 **(a)** n=20, p=0.25.$n = 20,\ p = 0.25.$ **(b)** 5. **(c)** P(X=5)=0.202.$P(X = 5) = 0.202.$

12.33 **(a)** Binomial is reasonable, n=30, p=0.135.$n = 30,\ p = 0.135.$ **(b)** μ=4.05.$\mu = 4.05.$ **(c)** 0.987. **(d)** 0.381.

12.35 **(a)** n=6, p=0.75.$n = 6,\ p = 0.75.$ **(b)** S={0,1,2,3,4,5,6}.$S = \{0, 1, 2, 3, 4, 5, 6\}.$ **(c)** P(X=0)=0.00024,$P(X = 0) = 0.00024,$ P(X=1)=0.0044,$P(X = 1) = 0.0044,$ P(X=2)=0.03296,$P(X = 2) = 0.03296,$ P(X=3)=0.1318,$P(X = 3) = 0.1318,$ P(X=4)=0.2966,$P(X = 4) = 0.2966,$ P(X=5)=0.3560,$P(X = 5) = 0.3560,$ P(X=6)=0.1780.$P(X = 6) = 0.1780.$ **(d)** μ=4.5, σ=1.06.$\mu = 4.5,\ \sigma = 1.06.$

12.37 **(a)** S={0,1,2,3,4,5}.$S = \{0, 1, 2, 3, 4, 5\}.$ **(b)** P(X=0)=0.2373,$P(X = 0) = 0.2373,$ P(X=1)=0.3955,$P(X = 1) = 0.3955,$ P(X=2)=0.2637, P(X=3)=0.0879, P(X=4)=0.0146, P(X=5)=0.0010.$P(X = 2) = 0.2637,\ P(X = 3) = 0.0879,\ P(X = 4) = 0.0146,\ P(X = 5) = 0.0010.$ **(c)** μ=1.25, σ=0.968.$\mu = 1.25,\ \sigma = 0.968.$

12.39 The probability is nearly 1.

12.41 **(a)** *P*(1 of 17 vaccinated children) = 0.3741. *P* (1 of 3 unvaccinated children) = 0.0960. *P*(1 in each group) = 0.0359. **(b)** Total probability is 0.1977.

12.43 **(a)** Poisson. μ=3.6,$\mu = 3.6,$ σ=1.897.$\sigma = 1.897.$ **(b)** 0.0273 and 0.9727, respectively.

12.45 P(X≥5)=0.219, P(X≥6)=0.105, P(X≥7)=0.045. $P(X \geq 5) = 0.219,\ P(X \geq 6) = 0.105,\ P(X \geq 7) = 0.045.$ Reserving 6 beds for ER admissions would be enough 95.5% of the time.

12.47 **(a)** Poisson. μ=15.58,$\mu = 15.58,$ σ=3.947.$\sigma = 3.947.$ **(b)** P(X=0)= $P(X = 0) =$ 1.71×10−7,1.71×10^{-7}, P(X≤5)=$P(X \leq 5) =$ 0.00186,$0.00186,$ P(X≤15)=0.509, P(X≤25)=0.990, P(X>25)=0.01. $P(X \leq 15) = 0.509,\ P(X \leq 25) = 0.990,\ P(X > 25) = 0.01.$ **(c)** P(X≥48)=3.53×10−11.$P(X \geq 48) = 3.53 \times 10^{-11}.$ This is too unlikely to be due to random, isolated cases. It points to an epidemic. The contaminated food must have been distributed at least in South Dakota and Wisconsin.

12.49 **(a)** P(X>10)=5.5×10−7.$P(X > 10) = 5.5 \times 10^{-7}.$ **(b)** Zero. **(c)** Zero. **(d)** Yes.

12.51 **(a)** 0.0651. **(b)** 0.0605. **(c)** Fourth=0.0563,$\text{Fourth} = 0.0563,$ fifth=0.0524,$\text{fifth} = 0.0524,$ kth=(0.07)(0.93)k−1.$k\text{th} = (0.07)(0.93)^{k-1}.$

12.53 **(a)** 0.12705. **(b)** 0.12705, 0.12705. **(c)** 0.38115.

Chapter 13

13.1 1.3 is a parameter; 1.75 is a statistic.

13.3 **(a)** 130.2. **(b)** Sampled 1, 4, 5, 9. Values=99,125,170,147; $\text{Values} = 99, 125, 170, 147;$ x̄=135.25.$\bar{x} = 135.25.$ **(c)** x̄$\bar{x}$ repeated 9 times: 119.5, 125.75, 141.25, 138, 118, 119, 137.75, 123, 132.5. The center of all these is about 129. **(d)** There are way too many numbers to do by hand.

13.5 **(a)** x̄$\bar{x}$ is not systematically higher or lower than μ.$\mu.$ **(b)** With large samples, x̄$\bar{x}$ is more likely to be close to μ.$\mu.$

13.7 **(a)** *N*(188, 4.1). 0.5346. **(b)** 0.9792.

13.9 **(a)**Strongly right-skewed. **(b)** Approximately Normal. **(c)** Somewhat right-skewed, but not Normal.

13.11 **(a)** 6% is a parameter, 4% is a statistic. **(b)** Mean=0.06,$\text{Mean} = 0.06,$ standard deviation=0.0194.$\text{standard deviation} = 0.0194.$

13.13 0.1512.

13.15 If 1 of the 12 workers had a serious injury, it would cost more than the premiums collected. For thousands of policies, the average claim should be close to $439.

13.17 (c)

13.18 (a)

13.19 (a)

13.20 (c)

13.21 (a)

13.22 (c)

13.23 (b)

13.24 (c)

13.25 (b)

13.26 (b)

13.27 65 inches is a statistic; 64 inches is a parameter.

13.29 **(a)** 0.2812. **(b)** Standard deviation=2.8⁄n.$\text{Standard deviation} = 2.8/\sqrt{n}$. Required sample size=32.$\text{Required sample size} = 32$. **(c)** 0.9566.

13.31 **(a)** 1/6. **(b)** 283.

13.33 **(a)** 0.067. **(b)** 0.0013.

13.35 About 3.64.

13.37 **(a)** Mean=79.8 years,$\text{Mean} = 79.8\,\text{years}$, standard deviation=1.55 years.$\text{standard deviation} = 1.55\,\text{years}$. **(b)** Central limit theorem (n=100).$(n = 100)$. **(c)** 0.1228.

13.39 Given the very strong right-skew, n≥40$n \geq 40$ would be enough to be able to apply the central limit theorem.

13.41 For all practical purposes, it is a parameter. The census tracks everyone (except homeless individuals).

13.43 **(a)** The mean proportion of Hispanics is 0.139. The standard deviation is 0.0089. **(b)** P(p^≤0.12)=0.0167.$P(\hat{p} \leq 0.12) = 0.0167$.

13.45 **(a)** Mean=0.05,$\text{Mean} = 0.05$, standard deviation=0.05×0.95⁄n.$\text{standard deviation} = \sqrt{0.05 \times 0.95/n}$. **(b)** P(p^≥0.07)=$P(\hat{p} \geq 0.07) =$ 0.0972.0.0972. **(c)** P(p^≥0.08)=0.0258, P(p^≥0.10)=0.0006.$P(\hat{p} \geq 0.08) = 0.0258,\ P(\hat{p} \geq 0.10) = 0.0006$.

Chapter 14

14.1 **(a)** 0.4. **(b)** 0.8 cm. **(d)** 95%.

14.3 **(a)** (144, 158) dollars. **(b)** Parameter: mean weekly spending on food *μ* for all American adults. The procedure results in a 95% chance of capturing the true value μ.μ.

14.5 We are 95% confident that the concentration of active ingredient in this batch is between 0.8327 and 0.8481 g/l.

14.7 The parameter is μ,μ, the mean hemoglobin level in Jordan's children population. H0: μ=12, Ha: μ<12.$H_0\colon \mu = 12,\ H_a\colon \mu < 12.$

14.9 **(a)** *N*(12 g/dl, 0.2263 g/dl). **(b)** P=0.0010,$P = 0.0010,$ reject H0.H_0. **(c)** P=0.1884,$P = 0.1884,$ fail to reject H0.H_0. **(d)** The first *P*-value is very low, which says that the study outcome would be very unlikely if H0H_0 was true (so we reject H0H_0). The second *P*-value is not low, which is inconclusive (so we fail to reject H0H_0). **(e)** The first *P*-value is statistically significant at the α=0.05$\alpha = 0.05$ level, but the second is not. However, the actual *P*-values are more informative, because we can conclude that the first study provided *very strong* evidence against H0H_0 (P=0.001).$(P = 0.001)$.

14.11 **(a)** A response at least as large at that observed would be seen almost half the time (just by chance) even if there was no effect in the entire population of rats. **(b)** The differences observed were small enough that they would often happen just by chance if there was no effect in the entire population of rats.

14.13 Let *μ* be the mean IQ for all seventh-grade girls in this district. We test H0: μ=100, Ha: μ≠100$H_0\colon \mu = 100,\ H_a\colon \mu \neq 100$ (two-sided because we have no prior theory why it should be a directional effect) and find z=2.17, P=0.0302.$z = 2.17,\ P = 0.0302.$ There is reasonably strong evidence that the mean IQ differs from 100 (it is greater) in this population.

14.15 **(a)** Yes; 10 is within the 95% confidence interval, because the *P*-value is not low enough to reject at the α=0.05$\alpha = 0.05$ level. **(b)** No; 10 is not in the 90% confidence interval, because the *P*-value is low enough to reject at the α=0.10$\alpha = 0.10$ level.

14.17 **(a)** No. **(b)** Yes.

14.18 (c)

14.19 (a)

14.20 (c)

14.21 (b)

14.22 (b)

14.23 (a)

14.24 (b)

14.25 (b)

14.26 (b)

14.27 (a)

14.29 The margin of error says that 95% of surveys should contain the true population proportion, not necessarily the proportion obtained by this survey.

14.31 **(a)** 2.138. **(b)** 1.96. **(c)** 91.1±(1.96)(2.138).$91.1 \pm (1.96)(2.138)$. **(d)** Smaller, because the critical value would be smaller (1.645).

14.33 **(a)** We are 99% confident that the mean baseline LDL level in the population of adult with diabetes is between 123.1 and 128.1 mg/dl. **(b)** LDL levels may be different in the general and diabetic adult populations in terms of both center (μ)(μ) and variability (σ).(σ).

14.35 μ is the mean daily sodium intake (in mg) in the American teenage population. H0: μ=2300, Ha: μ>2300.$H_0\colon \mu = 2300,\ H_a\colon \mu > 2300$.

14.37 **(a)** She had no expectations before she looked at the data. She cannot change her hypothesis just to fit the data. **(b)** P=0.0930. $P = 0.0930$.

14.39 It is essentially correct.

14.41 Something that occurs “less than once in 100 repetitions” also occurs “less than 5 times in 100 repetitions,” but not vice versa.

14.43 **(a)** μ is the mean perceived gender-orientation of cinnamon in the American adult population. H0: μ=12, Ha: μ≠12.$H_0\colon \mu = 12,\ H_a\colon \mu \neq 12$. **(b)** A non-significant *P*-value is inconclusive (H0H_0 may or may not be true). **(c)** The study was inconclusive and failed to find evidence that cinnamon doesn’t have a gender-neutral smell (*P* not statistically significant).

14.45 It is possible to determine statistical significance using both *P*-values and confidence intervals. However, *P*-values only compare the parameter to a hypothesis, while confidence intervals actually give a range of estimates for the parameter.

14.47 **(a)** Yes (*P*-value very close to zero). **(b)** 10.0021 to 10.0025. The scale is slightly biased on the heavy side. **(c)** *P*-values tell us only that the scale is off, whereas confidence intervals tell us how much it is off (not too much, in fact): Confidence intervals are more informative.

Chapter 15

15.1 **(a)** No; the poll is a voluntary survey. **(b)** No; inference is still inappropriate, because the poll is a voluntary survey.

15.3 There might be a pattern to these consecutive records, and they come from only one academic medical center—they may not be representative of other procedures performed elsewhere.

15.5 **(a)** 3.92. **(b)** 1.96 (for $n = 100$) and 0.98 (for $n = 400$). **(c)** As the sample size increases, the margin of error decreases.

15.7 **(a)** n=5, P=0.1855;$n = 5,\ P = 0.1855;$ n=15,$n = 15,$ P=0.0607;$P = 0.0607;$ n=40,$n = 40,$ P=0.0057.$P = 0.0057.$ **(b)** As the sample size increases, the sampling distribution gets narrower. **(c)** Effect size (an estimate of the true value of μ) rather than simply whether μ differs from a reference value.

15.9 **(a)** H0:H_0: patient is healthy, Ha:H_a: patient is ill. Type I error: sending a healthy patient to the doctor. Type II error: clearing a patient who is ill.

15.11 **(a)** No; since we are running the test so many times, we would expect about 5 individuals to get $P \leq 0.01$. **(b)** Re-administer the ESP test to these four individuals to see if their performance is confirmed.

15.13 The earlier studies, when the scientific consensus was that bacteria could not thrive in extremely low pH environment.

15.15 **(a)** 178. **(b)** 307. A higher confidence level requires a larger sample size.

15.17 **(a)** *P*(Type I error) = 0.05; *P*(Type II error) = 0.22. **(b)** 0.86.

15.18 (a)

15.19 (c)

15.20 (b)

15.21 (c)

15.22 (b)

15.23 (a)

15.24 (a)

15.25 (a)

15.26 (b)

15.27 (c)

15.29 People are likely to lie about their sex lives. The margin of error does not take this into account.

15.31 B, A, C, respectively.

15.33 **(a)** Statistical significance is not the same as practical significance. **(b)** A significance test performed with a large sample size can be statistically significant even if the effect is small (high power).

15.35 **(a)** The margin of error decreases. **(b)** The *P*-value gets smaller. **(c)** The outlier will have a greater effect on a small sample.

15.37 **(a)** 166. **(b)** 97.

15.39 **(a)** There are not enough data to determine whether the outcome was due to chance under the null hypothesis. **(b)** Small sample size or large variability.

15.41 Power 0.381, 0.751, and 0.984. As the sample size increases, power increases. For a test that almost always rejects H0H_0 when μ=4.7, $\mu = 4.7,$ we should use n=40.$n = 40.$

15.43 **(a)** H0: μ=128, Ha: μ≠128.H_0: $\mu = 128$, H_a: $\mu \neq 128.$ **(b)** Power=0.924. $\text{Power} = 0.924.$ **(c)** Power=0.924.$\text{Power} = 0.924.$ Yes. **(d)** The power would be higher.

Chapter 16

16.1 **(a)** All probabilities are between 0 and 1 and they add up to 1. **(b)** A newborn is a woman’s first or second child (at most her second child): $P(X \le 2)=0.710$.$P(X \le 2) = 0.710.$ A newborn is a woman’s sixth or later child: $P(X>5)=0.022$.$P(X > 5) = 0.022.$ **(c)** The event $X=3$,$X = 3,$ which has probability 0.168.

16.3 **(a)** Mean=2.108,$\text{Mean} = 2.108,$ standard deviation=0.1259. $\text{standard deviation} = 0.1259.$ **(b)** Approximately Normal. **(c)** 0.195.

16.5 **(a)** It is on or above the horizontal axis everywhere, and the total area beneath the curve is 1. **(b)** 1/3. **(c)** 2/3. **(d)** 1/3.

16.7 **(a)** 0.5899. **(b)** 384.2 mg.

16.9 400. We expect 400 individual weights to be as variable as 100 individual weights (only sample *averages* are less variable with larger sample sizes).

16.11 **(a)** H0: μ=188,$H_0\colon \mu = 188,$ Ha: μ<188, z=−1.46, P=0.0721; $H_a\colon \mu < 188,\ z = -1.46,\ P = 0.0721;$ significant only at the α=0.10$\alpha = 0.10$ level. **(b)** The value 188 doesn’t fall within the confidence interval; therefore, the results are not significant at the α=0.05$\alpha = 0.05$ level for the one-sided alternative.

16.13 84.8 to 90.4.

16.15 **(a)** The treatment was not assigned. **(b)** Socioeconomic and hereditary factors may also affect IQ.

16.17 Answers may vary because some points cover a spectrum of issues. Mathematical meaning: 1, 2. Data collection and interpretation: 4, 5. Both concerns: 3, 6.

16.19 **(a)** In each group, all probabilities are between 0 and 1 and they add up to 1. **(b)** X<3$X < 3$ is “intense physical activity on fewer than 3 days last week.” P(X<3 | girl)=0.43.$P(X < 3 \mid \text{girl}) = 0.43.$ P(X<3 | boy)=0.31.$P(X < 3 \mid \text{boy}) = 0.31.$ **(c)** This is X≥1.$X \ge 1.$ P(X≥1 | boy)=0.92.$P(X \ge 1 \mid \text{boy}) = 0.92.$ **(d)** They are dependent because, for example, P(X<3 | girl)≠P(X<3 | boy).$P(X < 3 \mid \text{girl}) \ne P(X < 3 \mid \text{boy}).$

16.21 **(a)** 0.0746 for girls; 0.2496 for boys. **(b)** 0.4013 for girls; 0.7516 for boys. **(c)** 0.5 girls; 1.3 boys. **(d)** These are expectations, on average, over very many random samples of 10 girls or 10 boys.

16.23 **(a)** *P*(MAD | DA) = 2 × *P*(MAD). **(b)** A person can have both MAD and DA, so these events are not disjoint. *P*(MAD | DA) ≠

P(MAD), so these events are not independent (the probability of MAD depends on DA status).

16.25 **(a)** Binomial, n=5,$n = 5,$ p=0.5.$p = 0.5.$ **(b)** *P*(5 boys) = 0.03125, *P*(4 boys, 1 girl) = 0.15625. **(c)** *P*(1 girl | 4 boys) = 0.5. This is a conditional probability.

16.29 **(a)** There aren't enough data to show that the outcome was very unlikely by chance alone. **(b)** It is a relatively small effect size. This would explain why previous results failed to reject the null hypothesis. **(c)** The older siblings in that age range most likely have more education and therefore a slightly higher IQ score.

Chapter 17

17.1 s=3.79.$s = 3.79.$

17.3 **(a)** 2.518. **(b)** 2.064. **(c)** 0.0150.

17.5 569.7 to 698.3 cm^2.

17.7 The data are roughly symmetric, with x̄=11.80 mg,$\bar{x} = 11.80\ \mathrm{mg},$ s=1.18 mg,$s = 1.18\ \mathrm{mg},$ n=92.$n = 92.$ The 95% confidence interval for the mean forewing weight of monarch butterflies is 11.55 to 12.04 mg.

17.9 Data are strongly skewed to the left, but the sample size is relatively large (n=29),$(n = 29)$, so the test should be relatively accurate; P=4.2×10−10,$P = 4.2 \times 10^{-10}$, extremely significant.

17.11 **(a)** H0: μ=500, Ha: μ≠500.$H_0\colon \mu = 500,\ H_a\colon \mu \neq 500.$ **(b)** There is weak evidence (t=−1.93,$(t = -1.93,$ P=0.064)$P = 0.064)$ that the Fitbit Flex is inaccurate, on average, for 500 steps walked. **(c)** There is strong evidence (t=2.98,$(t = 2.98,$ P=0.006)$P = 0.006)$ that the iPhone 5s Moves app is inaccurate, on average, for 500 steps walked.

17.13 **(a)** The sample sizes are identical, the magnitude of the effect (the absolute value of x̄-μ)$\bar{x} - \mu)$ is fairly similar for both trackers, but the Fitbit Flex was much more variable (s=93$(s = 93$ compared with s=55$s = 55$ for the app). We can never conclude about effect size just from a *P*-value because the *P*-value depends on several factors (effect size, variability, and sample size are often the most relevant). **(b)** Each test comes with the possibility of making a type

1 error. Multiple tests increase the chance of making at least one type 1 error in the set.

17.15 **(a)** We are 95% confident that, in the population of subjects diagnosed with essential tremor, the mean difference in tremor amplitude without and with ACT is between 0.82 and 1.33 cm; that is, tremor amplitude in this population is lower by 0.82 to 1.33 cm, on average, when the ACT is turned on. **(b)** For the 11 pairwise differences, $\bar{x}$=1.073 cm,$\bar{x} = 1.073\text{ cm},$ s=0.380 cm.$s = 0.380\text{ cm}.$

17.17 **(a)** They should be representative of pregnant women in this geographical area, but they could have very different diets and lifestyles than pregnant women elsewhere. **(b)** The *t* procedures are robust to deviations from Normality when the sample size is large (in this case n=4134,$n = 4134,$ very large). **(c)** Part (a)—the interpretation would need to specify this particular population carefully.

17.19 (b)

17.20 (b)

17.21 (c)

17.22 (a)

17.23 (c)

17.24 (a)

17.25 (c)

17.26 (b)

17.27 (a)

17.26 (c)

17.29 **(a)** $\bar{x}$=15.59 feet,$\bar{x} = 15.59\text{ feet},$ s=2.550 feet.$s = 2.550\text{ feet}.$ **(b)** 14.81 to 16.37 feet. We reject the claim that μ=20 feet.$\mu = 20\text{ feet}.$ **(c)** What population are we examining? Full-grown sharks? Male sharks? Can this be considered an SRS?

17.31 10.04±0.49.$10.04 \pm 0.49.$

17.33 The data are slightly skewed to the right, but the sample is large enough for the 95% *t* confidence interval.

17.35 **(a)** Only if the students were randomly chosen from the population of all male college students. No major departure from Normality.

(b) 425.4 ± 32.4.

17.37 **(a)** A dotplot shows no major causes for concern. **(b)** 0.92 to 1.42 days.

17.39 **(a)** 0.2450 to 0.3113 g/m^2. **(b)** Yes, because the sample size is large enough.

17.41 **(a)** A subject’s responses to the two treatments would not be independent. **(b)** Yes (t=−4.41, P=0.0069).$(t = -4.41,\ P = 0.0069).$

17.43 This is a matched pairs setting; compute the differences for collaboration required minus collaboration not required. H0: μ=0, $H_0\colon \mu = 0,$ Ha: μ>0,$H_a\colon \mu > 0,$ t=7.37,$t = 7.37,$ df=7,$\text{df} = 7,$ P<0.0005. $P < 0.0005.$

17.45 H0: μ=1, Ha: μ>1,$H_0\colon \mu = 1,\ H_a\colon \mu > 1,$ t=8.402,$t = 8.402,$ P=0.0002. $P = 0.0002.$ However, the data may be skewed to the left, in which case the sample size would not be large enough for inference.

Chapter 18

18.1 Blind spot size. Matched pairs.

18.3 Concentration. Single sample.

18.5 **(a)** For the HSAM group, x̄=6.00, s=1.63,$\bar{x} = 6.00,\ s = 1.63,$ SE=0.52=1.63⁄10.$\text{SE} = 0.52 = 1.63/\sqrt{10}.$ For the control group, x̄=3.61, s=1.97,$\bar{x} = 3.61,\ s = 1.97,$ SE=0.47=1.97⁄18.$\text{SE} = 0.47 = 1.97/\sqrt{18}.$ t=(6.00−3.61)⁄$t = (6.00 - 3.61)/$ (0.522+0.472)=3.44. $(\sqrt{0.52^2 + 0.47^2}) = 3.44.$ **(b)** There is significant evidence (t=3.44, P=0.001,$(t = 3.44,\ P = 0.001,$ df=21)$\text{df} = 21)$ that individuals with HSAM have higher visual memory test scores than typical individuals, on average.

18.7 **(a)** A standard *t* distribution is always wider than the standard Normal distribution. So, we can tell right away that the value t= −0.771$t = -0.771$ is so small that it would give a large, nonsignificant *P*-value. **(b)** Echinacea: n=52,x̄=13.21,$n = 52, \bar{x} = 13.21,$ s=1.91×52=13.77.$s = 1.91 \times \sqrt{52} = 13.77.$ Placebo: n=103,$n = 103,$ x̄=15.05,$\bar{x} = 15.05,$ s=1.43×103=14.51.$s = 1.43 \times \sqrt{103} = 14.51.$ Software gives t=−0.771,$t = -0.771,$ df=107.3,$\text{df} = 107.3,$ P=0.22$P = 0.22$ (one sided).

18.9 **(a)** The two populations are all male mice susceptible to Zika, those infected and those not infected with Zika. The response variable is weight of the testes (in grams). H0: μinf.=μuninf.,$H_0\colon \mu_{\text{inf.}} = \mu_{\text{uninf.}}$, Ha: μinf.≠μuninf.$H_a\colon \mu_{\text{inf.}} \neq \mu_{\text{uninf.}}$ **(b)** There is an extreme outlier and the total sample size is fairly small (14).

18.11 (c)

18.12 (b)

18.13 (b)

18.14 (c)

18.15 (b)

18.16 (c)

18.17 (a)

18.18 (b)

18.19 (b)

18.20 (c)

18.21 **(a)** The female data are a bit skewed, but the combined sample size is large enough. **(b)** 10.4 to 22.9 mm (df=32).$(\text{df} = 32)$.

18.23 **(b)** H0: μG=μB,Ha: μG≠μB, t=1.64, P=0.1057, $H_0\colon \mu_G = \mu_B, H_a\colon \mu_G \neq \mu_B, t = 1.64, P = 0.1057,$ not statistically significant.

18.25 **(a)** Plotting the data reveals no major causes for concern. **(b)** t=2.57,$t = 2.57,$ df=24,$\text{df} = 24,$ P=0.008$P = 0.008$ (one-sided), statistically significant.

18.27 **(a)** The diabetic potentials appear to be large. **(b)** H0: μD=μN,Ha: μD≠μN,$H_0\colon \mu_D = \mu_N, H_a\colon \mu_D \neq \mu_N,$ t=3.077,$t = 3.077,$ P=0.0039 (df=36.60).$P = 0.0039\ (\text{df} = 36.60)$. **(c)** t=3.841,$t = 3.841,$ P=0.0005 (df=37.15).$P = 0.0005\ (\text{df} = 37.15)$.

18.29 H0: μ1=μ2,Ha: μ1<μ2, t=−10.29,$H_0\colon \mu_1 = \mu_2, H_a\colon \mu_1 < \mu_2, t = -10.29,$ P (df=36.89)=1.07×10−12.$P\ (\text{df} = 36.89) = 1.07 \times 10^{-12}$.

18.31 **(a)** H0: μTP=μno,Ha: μTP>μno,$H_0\colon \mu_{TP} = \mu_{no}, H_a\colon \mu_{TP} > \mu_{no},$ t=5.09, $t = 5.09,$ df=57,$\text{df} = 57,$ P=0.000002<0.01,$P = 0.000002 < 0.01,$ and highly statistically significant. There is very strong evidence that young adult males with tongue piercing have more enamel cracks, on average, than similar individuals with no tongue piercing. **(b)** No;

this is an observational study, so causation cannot be established.

18.33 (−2.9,9.7).$(-2.9, 9.7)$. The evidence is inconclusive on the effectiveness of delayed cord clamping.

18.35 **(a)** H0: μa=μw,$H_0\colon \mu_a = \mu_w$, Ha: μa<μw,$H_a\colon \mu_a < \mu_w$, t=−2.31, P (df=106.75)=0.0115.$t = -2.31,\ P\,(\mathrm{df} = 106.75) = 0.0115$. On average, people receiving acupuncture had fewer migraines than those on the wait list. **(b)** 1.2 ± 1.03. On the low end the reduction is almost nothing, but on the high end it is about 50%.

18.37 If those who had fewer migraines dropped out the most, then those remaining would have a higher migraine rate. Since the wait list had the highest dropout rates, this would imply that the wait list group would have a disproportional increase in migraines. If those who believed least in the acupuncture dropped out, the remaining subjects might feel that only acupuncture can help them.

18.39 **(a)** Two-sample *t* test, P=0.992.$P = 0.992$. **(b)** Compute the weight difference for each group, then find a two-sample confidence interval: 11.42 ± 6.88 with df rounded to 17. The control group gained significantly more weight than the group receiving ink applications. The ink appears to be toxic.

18.41 **(a)** H0: μm=μp,Ha: μm≠μp,$H_0\colon \mu_m = \mu_p, H_a\colon \mu_m \neq \mu_p$, t=0.4594, P (df=39.6)=0.6484.$t = 0.4594,\ P\,(\mathrm{df} = 39.6) = 0.6484$. **(b)** Were the data skewed or bimodal? Were there any outliers?

Chapter 19

19.1 **(a)** The population was cultures from individuals diagnosed with strep. *p* = proportion resistant to antibiotic in the population. **(b)** p^=0.568.$\hat{p} = 0.568$.

19.3 This online poll is a voluntary response sample and cannot be considered an SRS of any larger population.

19.5 p^=0.5677. 0.544 to 0.591.$\hat{p} = 0.5677.\ 0.544 \text{ to } 0.591$.

19.7 The lowest count, 11, is too small for the large-sample method.

19.9 **(a)** p≅0.5675. 0.544 to 0.591.$\tilde{p} = 0.5675.\ 0.544 \text{ to } 0.591$. **(b)** Both methods give similar results when the sample size is large because the impact of the extra 4 data points is negligible within the large sample.

19.11 1537.

19.13 H0: p=0.5, Ha: p>0.5, z=0.21, P=0.4174, $H_0: p = 0.5,\ H_a: p > 0.5,\ z = 0.21,\ P = 0.4174,$ not statistically significant.

19.15 **(a)** np0=2<10.$np_0 = 2 < 10.$ **(b)** n(1-p0)=8<10.$n\,(1 - p_0) = 8 < 10.$

19.16 (c)

19.17 (a)

19.18 (a)

19.19 (b)

19.20 (b)

19.21 (c)

19.22 (c)

19.23 (b)

19.24 (a)

19.25 (a)

19.27 **(a)** The number of successes is less than 15. However, n>10,$n > 10,$ so the plus four confidence interval is appropriate. **(b)** 0.125 ± 0.0636.

19.29 683.

19.31 0.593 to 0.612.

19.33 **(a)** H0: p=1⁄7, Ha: p>1⁄7.$H_0: p = 1/7,\ H_a: p > 1/7.$ **(b)** z=5.556,$z = 5.556,$ P<0.0001.$P < 0.0001.$ The dogs do significantly better than expected by chance alone.

19.35 0.2752 ± 0.0086. We are 95% confident that between 26.7% and 28.4% of all possible tap water specimens in D.C. would exceed the lead Action Level.

19.37 0.169 to 0.171. We are 95% confident that between 16.9% and 17.1% of American adults have ever been told by a physician or a nurse that they have depression.

19.39 The conditions for inference are met. We test H0: p=0.5, Ha: p>0.5 $H_0: p = 0.5,\ H_a: p > 0.5$ and find z=4.74, P=0.000001,$z = 4.74,\ P = 0.000001,$ highly significant. The plus four 95% confidence interval is 0.7328 to 0.9490. There is very strong evidence that young children prefer

Graham Crackers snacks packaged with popular cartoon images (P<0.001).$(P < 0.001)$. With 95% confidence we estimate that between 73.3% and 94.9% of young children show such a preference.

19.41 0.176 to 0.194. We are 95% confident that between 17.6% and 19.4% of all surgery patients are carriers of S.$S.$ *aureus*.

Chapter 20

20.1 p^abrupt=0.220, p^gradual=0.155. 0.0071 to 0.1224.
$\hat{p}_{\text{abrupt}} = 0.220$, $\hat{p}_{\text{gradual}} = 0.155$. 0.0071 to 0.1224.

20.3 **(a)** Not enough of the subjects taking echinacea failed to develop a cold. **(b)** 0.0524 ± 0.1079 for pechinacea-pplacebo.$p_{\text{echinacea}} - p_{\text{placebo}}$.

20.5 **(a)** Infants with severe respiratory distress. Randomized experiment. Treatment provided (explanatory) and infant survival (response). **(b)** H0: pbCPAP=pcontrol,H_0: $p_{\text{bCPAP}} = p_{\text{control}}$, Ha: pbCPAP>pcontrol$H_a$: $p_{\text{bCPAP}} > p_{\text{control}}$ **(c)** Yes. The data are reasonably random and representative of the corresponding populations. The counts of successes and failures in both samples are large enough. **(d)** z=2.36, P=0.0091.$z = 2.36$, $P = 0.0091$. The study provides significant evidence that the survival rate is greater with bCPAP than with nasal oxygen therapy in the population of Malawi infants with severe respiratory distress.

20.7 **(a)** p^asp=0.000906, p^plac=0.002356.$\hat{p}_{asp} = 0.000906$, $\hat{p}_{plac} = 0.002356$. **(b)** RR=0.3845.$RR = 0.3845$. Individuals in the aspirin group were 38% less likely to suffer a fatal heart attack compared with individuals in the placebo group. **(c)** OR=0.3840.$OR = 0.3840$. In this case, *OR* and *RR* are very close.

20.9 **(a)** ARR=0.00145,$ARR = 0.00145$, RRR=0.6155.$RRR = 0.6155$. *ARR* is the difference in risk between treatment and control; *RRR* is the reduction in risk in the treatment group relative to the risk of the control group. **(b)** 690. This means that 690 physicians would need to take aspirin daily for 5 years to save 1 of them from a fatal heart attack compared with taking a placebo daily for 5 years.

20.11 (a)

20.12 (c)

20.13 (a)

20.14 (c)

20.15 (c)

20.16 (b)

20.17 (b)

20.18 (b)

20.19 (b)

20.20 (c)

20.21 **(a)** There were no "successes" in the control group. **(b)** Treatment group: n=35$n = 35$ with 24 mice developing tumors. Control group: n=20$n = 20$ with 1 mouse developing a tumor. 0.636±0.238. $0.636 \pm 0.238.$ **(c)** The difference is significant at the 1% level because the 99% confidence interval (0.398 to 0.874) does not include zero.

20.23 **(a)** H0: pAZT=pplac,Ha: pAZT<pplac.$H_0\colon p_{\text{AZT}} = p_{\text{plac}}, H_a\colon p_{\text{AZT}} < p_{\text{plac}}.$ There are enough patients with and without AIDS to use a *z* procedure for this test. **(b)** P=0.0017.$P = 0.0017.$ **(c)** This means that neither the doctors nor the patients knew who had received the drug and who had received the placebo.

20.25 **(a)** Independent random samples and large enough counts of successes and failures in both samples, so the conditions are met. H0: ptreat=psham,$H_0\colon p_{\text{treat}} = p_{\text{sham}},$ Ha: ptreat>psham.$H_a\colon p_{\text{treat}} > p_{\text{sham}}.$ **(b)** z=0.98, P=0.1644.$z = 0.98,\ P = 0.1644.$ **(c)** The study failed to show that arthroscopic surgery does better than a sham procedure at treating patients with knee pain (P=0.1644).$(P = 0.1644).$

20.27 **(a)** H0: ptreat=pplac,Ha: ptreat<pplac.$H_0\colon p_{\text{treat}} = p_{\text{plac}}, H_a\colon p_{\text{treat}} < p_{\text{plac}}.$ z=−2.93, P=0.0017,$z = -2.93,\ P = 0.0017,$ statistically significant. **(b)** −0.075 to−0.013$-0.075 \text{ to } -0.013$ (plus four) or −0.074 to−0.014 $-0.074 \text{ to } -0.014$ (large sample).

20.29 The two samples are not independent, because the same 40 children were used for both conditions.

20.31 H0: pn=pf,Ha: pn<pf, z=−2.398, P=0.0082, $H_0\colon p_n = p_f, H_a\colon p_n < p_f,\ z = -2.398,\ P = 0.0082,$ statistically significant.

20.33 z=1.59, P=0.1118,$z = 1.59,\ P = 0.1118,$ not statistically significant.

20.35 **(a)** z=2.462, P=0.0069,$z = 2.462,\ P = 0.0069,$ statistically significant. **(b)** 0.331 to 0.663.

20.37 **(a)** Both parties will know whether liquid nitrogen or duct tape was applied. **(b)** A plus four 95% confidence interval is the only valid test: 0.2288±0.2334.$0.2288 \pm 0.2334.$ Because the interval (−0.005 to 0.462) contains the value zero, we fail to find enough evidence of a different outcome with the treatments (at a significance level of 0.05).

20.39 H0: pm=pf,Ha: pm≠pf,$H_0\colon p_m = p_f, H_a\colon p_m \neq p_f,$ z=0.5258,$z = 0.5258,$ P=0.599,$P = 0.599,$ no significant difference.

20.41 **(a)** z=−2.79, P=0.0052,$z = -2.79,\ P = 0.0052,$ significant. **(b)** z=2.16, P=0.0308,$z = 2.16,\ P = 0.0308,$ significant. **(c)** For women in their 40s, the proportion with increased LH levels is significantly greater in the placebo group. For women in their 50s and 60s, the proportion with increased LH levels is significantly greater in the melatonin group. The effect is reversed in the two age groups. **(d)** z=−0.68, P=0.4988,$z = -0.68,\ P = 0.4988,$ not significant. Pooling the two groups cancels out the two separate effects.

20.43 **(a)** OR=0.4249,$OR = 0.4249,$ RR=0.4474.$RR = 0.4474.$ **(b)** ARR=0.0483,$ARR = 0.0483,$ RRR=0.5526.$RRR = 0.5526.$ **(c)** NNT=21. $NNT = 21.$

Chapter 21

21.1 **(a)** Window orientation of bird strike, 3 levels. **(b)** H0: p0=p20=p40=1⁄3, Ha: H0$H_0\colon p_0 = p_{20} = p_{40} = 1/3,\ H_a\colon H_0$ is not true.

21.3 Expected counts are all =53⁄3=17.67.X2=16.11. $= 53/3 = 17.67. X^2 = 16.11.$

21.5 There is very strong evidence (P=0.0003)$(P = 0.0003)$ that not all window angles are equally likely to produce accidental bird strikes.

21.7 The expected counts and chi-square components show that more birds than expected strike the vertical window and fewer than expected strike the 40-degree window.

21.9 Data fall into a multinomial setting; expected counts are all larger than 5.0.

21.11 **(a)** Yes; data fall into a multinomial setting; expected counts are all larger than 5.0. **(b)** No; data fall into a multinomial setting but one of only four expected counts is smaller than 5.0.

21.12 (c)

21.13 (c)

21.14 (a)

21.15 (b)

21.16 (b)

21.17 (a)

21.18 (b)

21.19 (c)

21.20 (b)

21.21 (c)

21.23 **(a)** Yes; counts are high enough. **(b)** H0: pYR=9⁄16,$H_0\colon p_{YR} = 9/16,$ pYW=3⁄16,$p_{YW} = 3/16,$ pGR=3⁄16,$p_{GR} = 3/16,$ pGW=1⁄16,$p_{GW} = 1/16,$ Ha: H0$H_a\colon H_0$ is not true. **(c)** 312.75, 104.25, 104.25, and 34.75. **(d)** P=0.9254.$P = 0.9254.$ The data are consistent with Mendel's second law: The study failed to provide evidence to the contrary.

21.25 **(a)** Pregnancy: 0 observed, 13.2 expected. No pregnancy: 87 observed, 73.8 expected. Cohort can be considered a random sample. Expected counts are large enough. X2=15.56,$X^2 = 15.56,$ df=1,$\text{df} = 1,$ P=0.00008,$P = 0.00008,$ highly significant. **(b)** Pregnancy: 6 observed, 7.1 expected. No pregnancy: 29 observed, 27.9 expected. Expected counts are large enough. X2=0.21,$X^2 = 0.21,$ df=1,$\text{df} = 1,$ P=0.6438,$P = 0.6438,$ not significant. **(c)** There is very strong evidence that Plan B prevents pregnancy when taken before ovulation (X2=15.56, P=0.00008)$(X^2 = 15.56,\ P = 0.00008)$ but not when taken after ovulation (X2=0.21, P=0.6438). $(X^2 = 0.21,\ P = 0.6438).$

21.27 H0: pB=pW=pL=1⁄3, Ha: H0$H_0\colon p_B = p_W = p_L = 1/3,\ H_a\colon H_0$ is not true. X2=57.18, P=3.8×10−13.$X^2 = 57.18,\ P = 3.8 \times 10^{-13}.$ There is highly statistically significant evidence that American adults who drink alcohol do not favor equally beer, wine, and liquor. The study found more than expected who favor beer and fewer than expected who favor liquor.

21.29 H0: H_0: Probabilities match U.S. percents, Ha: H0H_a: H_0 is not true. X2=1709.8, df=4,$X^2 = 1709.8$, df $= 4$, P≈0.$P \approx 0$. There is highly statistically significant evidence that the ethnic breakdown in the population targeted by this study does not match the U.S. ethnic breakdown.

21.31 H0: Values of piH_0: Values of p_i are given by the model of random breeding; Ha: H0H_a: H_0 is not true. X2=9.377, df=7, P=0.2267. $X^2 = 9.377$, df $= 7$, $P = 0.2267$. The data are consistent with the hypothesis that prairie dogs breed randomly.

21.33 H0: pw=12⁄16, pv=3⁄16, pt=1⁄16, Ha: H0 H_0: $p_w = 12/16$, $p_v = 3/16$, $p_t = 1/16$, H_a: H_0 is not true. X2=0.0980, $X^2 = 0.0980$, df=2,df $= 2$, P=0.952,$P = 0.952$, not statistically significant. The data are consistent with a dominant epistatic model.

21.35 H0: pF=1⁄4, pSF=1⁄2, pS=1⁄4, Ha: H0 H_0: $p_F = 1/4$, $p_{SF} = 1/2$, $p_S = 1/4$, H_a: H_0 is not true. X2=0.7204,$X^2 = 0.7204$, df=2,df $= 2$, P=0.6975.$P = 0.6975$. The data are consistent with a codominance model of inheritance.

Chapter 22

22.1 **(a)** From left to right, row by row: 21, 268, 289, 17, 278, 295, 3, 70, 73. **(b)** From left to right, row by row: 7.3%, 92.7%, 5.8%, 94.2%, 4.1%, 95.9%. Conditional distribution of abstinence, given treatment group. The outcomes are fairly similar, but nicotine-based treatments may be a little more effective than a placebo.

22.3 **(a)** H0: H_0: No relationship between treatment and abstinence outcome in the population of adult smokers who want to quit. Ha: H0H_a: H_0 is not true. **(b)** 3. Multiple tests increase the chance of committing at least one Type I error.

22.5 **(a)** 18.04=(289)(41)⁄657.$18.04 = (289)(41)/657$. 270.96=(289)(616)⁄657. $270.96 = (289)(616)/657$. 18.04+270.96=289$18.04 + 270.96 = 289$ **(b)** The observed and expected counts are fairly similar, although the nicotine e-cigarette group had a few more individuals with verified smoking abstinence than expected. **(c)** 0.48745=(21−18.08)2⁄18.08.

$0.48745 = (21 - 18.08)^2/18.08$. **(d)** X2=1.202.$X^2 = 1.202$.

22.7 **(a)** P=0.548.$P = 0.548$. **(b)** The test is not statistically significant, so the chi-square components could reflect just random variations due to the randomization of subjects to treatments. **(c)** The experiment failed to find significant evidence that, in the population of adult smokers who want to quit smoking, the treatments offered are associated with different smoking cessation outcomes (P=0.548). $(P = 0.548)$.

22.9 **(a)** Only 1 of the 6 expected cell counts is slightly less than 5.0 (4.56>1.0),$5.0\ (4.56 > 1.0)$, so the cell count requirements are met. **(b)** The study is a completely randomized experiment, therefore the 3 groups form 3 independent random samples. The larger population represented is all adult smokers who want to quit smoking.

22.11 **(a)** X2=46.98,$X^2 = 46.98$, P=7.2×10−12,$P = 7.2 \times 10^{-12}$, highly statistically significant. The evidence of an association is very strong. **(b)** The cell count requirements for the chi-square test are for expected counts, not observed counts. The smallest expected count is 7.0>5.0,$7.0 > 5.0$, so those conditions are met.

22.13 (c)

22.14 (c)

22.15 (b)

22.16 (b)

22.17 (b)

22.18 (a)

22.19 (b)

22.20 (a)

22.21 (c)

22.22 (b)

22.23 **(a)** By groups of increasing age, for the positive test: 31.07%, 16.50%, 9.32%, 17.48%, 15.92%, 9.71%. By groups of increasing age, for the negative test: 34.99%, 7.40%, 8.96%, 16.93%, 17.21%, 14.51%. **(b)** Yes: All expected cell counts are well over 5.0 and the study used a representative sample. **(c)** X2=40.554, P<0.0005 $X^2 = 40.554$, $P < 0.0005$ (rounded to 0.000 by Minitab). There is very strong evidence of an association between age group and HPV status. **(d)** The 20 to 25 age group is the largest contributor: They are much more likely to be HPV positive.

22.25 **(a)** 35.3% (avoid) and 10.6% (consume). **(b)** Smallest expected count 11.03>5.0.$11.03 > 5.0.$ **(c)** The experiment provides strong evidence (X2=8.28 and P=0.0040)$(X^2 = 8.28 \text{ and } P = 0.0040)$ of an association between peanut avoidance in young children with a peanut allergy and persisting peanut allergy at 60 months of age: Those avoiding peanuts are more likely to remain allergic.

22.27 **(a)** Stress=0.909,$\text{Stress} = 0.909,$ exercise=0.794,$\text{exercise} = 0.794,$ regular treatment=0.700.$\text{treatment} = 0.700.$ **(b)** Expected heart attacks: stress=6.78,$\text{stress} = 6.78,$ exercise=6.99,$\text{exercise} = 6.99,$ regular treatment=8.22.$\text{regular treatment} = 8.22.$ All cells have expected counts above 5. **(c)** X2=4.84 and P=0.0889.$X^2 = 4.84 \text{ and } P = 0.0889.$ This is not statistically significant at the 0.05 level.

22.29 X2=11.975,$X^2 = 11.975,$ df=3,$\text{df} = 3,$ P=0.007.$P = 0.007.$ There is a significant relationship between birth order and sex. Males are more represented among eggs that hatch first and second.

22.31 H0:H_0: No relationship between tongue piercing and the presence of enamel cracks. X2=7.39, df=1, P=0.0065, $X^2 = 7.39, \text{ df} = 1, \ P = 0.0065,$ significant. There is strong evidence that, in young adult males, enamel cracks are more common when the tongue is pierced (82.6%) than when it isn’t (56.5%). However, causation cannot be established because this is an observational study.

22.33 H0:H_0: No relationship between species and life stage (the two variables are independent). The smallest expected count is 49.51>5.0.$49.51 > 5.0.$ There is very strong evidence of a relationship between species and life stage in this forest (X2=420.311, $(X^2 = 420.311,$ P<0.005).$P < 0.005).$ The largest contributors to X2X^2 come from the cells for sapling RC and for dead RC: There are many more sapling western red cedar trees and far fewer dead ones than would be expected if H0H_0 was true, suggesting that the western red cedar species is a successional species displacing the other established species in this forest.

22.35 The chi-square test found no significant relationship between treatment type and death following a stroke during the two-year treatment period (P=0.701).$(P = 0.701)$. It appears that none of the active treatments would do better than a placebo in preventing stroke-related death.

22.37 Neither test was statistically significant (P=0.577 and P=0.3099,respectively).$(P = 0.577 \text{ and } P = 0.3099, \text{respectively})$. The echinacea extracts used as treatment or preventively do not appear to do better than a placebo at preventing the common cold or reducing the number of cold symptoms.

22.39 H0: H_0: No association between gender and tumor rate, X2=0.2765, df=1,$X^2 = 0.2765,\ \text{df} = 1,$ P=0.5990.$P = 0.5990$. Not statistically significant.

Chapter 23

23.1 **(a)** r=−0.9282.$r = -0.9282$. **(b)** The slope ββ is the average additional algal net growth for each unit increase in the number of grazers. We estimate b=−0.5286 and a=−1.30.$b = -0.5286 \text{ and } a = -1.30$. **(c)** Residuals: −0.071,−0.143,0.686,−0.486,−0.157,0.171. $-0.071, -0.143, 0.686, -0.486, -0.157, 0.171$. The estimated standard deviation is s=0.443.$s = 0.443$.

23.3 **(a)** t=−4.99.$t = -4.99$. **(b)** df=4, P=0.0038$\text{df} = 4,\ P = 0.0038$ (0.0025<P<0.005$(0.0025 < P < 0.005$ with Table C). There is strong evidence of a negative linear relationship at the population level.

23.5 **(a)** t=−4.99,$t = -4.99,$ df=4,0.0025<P<0.005$\text{df} = 4, 0.0025 < P < 0.005$ (Table C). **(b)** |r|=0.9282, n=6,0.0025<P<0.005 $|r| = 0.9282,\ n = 6, 0.0025 < P < 0.005$ (Table E).

23.7 −0.8227 to−0.2344$-0.8227 \text{ to } -0.2344$ net growth rate per grazer.

23.9 **(a)** “95% PI.” **(b)** “95% CI:” −2.88571±(2.776)(0.188532). $-2.88571 \pm (2.776)\,(0.188532)$. 2.776 is the critical value of *t* for df=4$\text{df} = 4$ and 95% confidence level.

23.11 The relationship is definitely not linear, so a linear regression analysis is not valid. **(b)** H0: β=0$H_0\colon \beta = 0$ versus Ha: β≠0.$H_a\colon \beta \neq 0$. P≈0.$P \approx 0$. **(c)** A curved pattern can still show some directionality.

Software will fit the best straight line it can even when the deviations from this linear pattern are clearly not random.

23.13 (c)

23.14 (c)

23.15 (c)

23.16 (a)

23.17 (a)

23.18 (b)

23.19 (b)

23.20 (a)

23.21 (b)

23.22 (c)

23.23 **(a)** b=−0.408.$b = -0.408.$ For every additional kilogram of nitrogen per hectare, the species richness measure decreases by 0.408, on average. **(b)** r2=0.55.$r^2 = 0.55.$ The linear relationship accounts for 55% of the observed variation in species richness. **(c)** H0: β=0 $H_0\colon \beta = 0$ versus Ha: β≠0$H_a\colon \beta \neq 0$ (or β<0).$(\text{ or } \beta < 0).$ The small *P* says that there is very strong evidence that the population slope ββ is negative—that is, that species richness does decrease as the amount of nitrogen deposited increases.

23.25 **(b)** r=0.6769.$r = 0.6769.$ **(c)** t=2.60, P=0.0316,$t = 2.60,\ P = 0.0316,$ significant.

23.27 Excel's 95% confidence interval for ββ is 0.4739 to 7.8856; this agrees (except for roundoff error) with b±t*SEb=4.1797±(2.306)(1.607).$b \pm t^* \mathrm{SE}_b = 4.1797 \pm (2.306)\,(1.607).$

23.29 **(a)** 12.997 to 14.165 cm^3. **(b)** 11.088 to 16.074 cm^3. **(c)** The prediction interval. Probably not.

23.31 **(a)** 12.520 to 13.071 cm^3. **(b)** 11.439 to 14.153 cm^3. **(c)** The prediction interval. The 3D system is more accurate than the 2D version but maybe still not accurate enough for practical use.

23.33 The scatterplot shows a positive association. The regression equation is ŷ=0.1523+8.1676x, and r2=0.606. $\hat{y} = 0.1523 + 8.1676x, \text{ and } r^2 = 0.606.$ The slope is significantly different from 0($t=13.25$, df=114, $P<0.001$).$0\ (t = 13.25,\ \mathrm{df} = 114,\ P < 0.001)$.

23.35 **(a)** There is considerably greater scatter for larger values of the explanatory variable. **(b)** The distribution of residuals looks reasonably Normal.

23.37 **(a)** Negative linear relationship. The regression equation is $\hat{y}=0.6471-0.0104x$.$\hat{y} = 0.6471 - 0.0104x$. **(b)** The observations are independent. The standard deviation appears roughly constant, although there are fewer observations for large BMIs. The distribution of residuals is irregular but not markedly non-Normal. **(c)** -0.0155 to -0.0053.-0.0155 to -0.0053.

23.39 0.0259 to 0.6434 nmol/l.

23.41 **(a)** $t=a/\mathrm{SE}_a=-0.0127/0.0126=-1.00$.$t = a/\mathrm{SE}_a = -0.0127/0.0126 = -1.00$. No significant evidence that the intercept is not zero ($P=0.332$).($P = 0.332$). **(b)** For df=14$\mathrm{df} = 14$ and a 95% confidence level, $t^*=2.145$,$t^* = 2.145$, so the confidence interval is $-0.0127\pm(2.145)(0.0126)=-0.0398$ to 0.0144.$-0.0127 \pm (2.145)\,(0.0126) = -0.0398$ to 0.0144. $\alpha=0$$\alpha = 0$ cannot be rejected because this interval contains the value 0.

Chapter 24

24.1 **(a)** Randomly allocate 36 flies to each of four groups, which receive varying dosages of caffeine; observe length of rest periods. **(b)** H0:H_0: All groups have the same mean rest period, Ha:H_a: At least one group has a different mean rest period. The significant *P*-value leads us to conclude that caffeine reduces the length of the rest period.

24.3 **(a)** When the three means are similar, *F* is very small and *P* is near 1. **(b)** Moving any mean increases *F* and decreases *P*.

24.5 **(a)** The data suggest that logging reduces the number of trees per plot and that recovery is slow. **(b)** The means show the same pattern. **(c)** $F=11.43$,$F = 11.43$, $P=0.000205$, H_0: $\mu_1=\mu_2=\mu_3$, Ha:$P = 0.000205,\ H_0\colon \mu_1 = \mu_2 = \mu_3,\ H_a$: Not all the means are the same. There is strong evidence that tree density is affected by logging history, on average.

24.7 The dotplots show no extreme outliers or strong skewness (given the small sample sizes). The ratio of standard deviations is about 1.16.

24.9 The sample standard deviations are much too different (the ratio is 10.2).

24.11 **(a)** Yes; the ratio of standard deviations is 1.8. **(b)** The data set is large (110 observations). **(c)** x̄=20.25,$\bar{x} = 20.25,$ MSG=4504.4. $\text{MSG} = 4504.4.$ MSE=1765.0.$\text{MSE} = 1765.0.$ **(d)** There is some weak evidence that the three exercise programs do not all yield the same mean subcutaneous fat reduction (F=2.552, P=0.0827); $(F = 2.552,\ P = 0.0827);$ however, this is not statistically significant at a typical significance level of 0.05.

24.12 (c)

24.13 (c)

24.14 (b)

24.15 (b)

24.16 (a)

24.17 (a)

24.18 (a)

24.19 (a)

24.20 (c)

24.21 (a)

24.23 Populations: normal-weight men, overweight men, obese men. Response variable: triglyceride level. H0: H_0: Men in different weight categories have the same triglyceride level on average, Ha: H_a: At least one weight group has a different mean triglyceride level. df=2 and 1821.$\text{df} = 2 \text{ and } 1821.$

24.25 **(a)** Yes, the men were recruited randomly and the sample standard deviations are very similar. **(b)** A *P*-value, no matter how small, says nothing about the size or importance of an effect. **(c)** The study found very strong evidence (P=0.000003)$(P = 0.000003)$ that, on average, triglyceride levels among Canadian men ages 34 to 64 depend on the men’s weight group.

24.27 **(a)** Yes; the mean control emission rate is half of the smallest of the others. **(b)** H0: H_0: All groups have the same mean emission rate, Ha: H_a: At least one group has a different mean emission rate. **(c)** Are the data Normally distributed? Were these random samples? **(d)** s=SEM×8.$s = \text{SEM} \times \sqrt{8}$. The rule-of-thumb ratio is 1.4755. **(e)** The study is inconclusive; it failed to find statistically significant evidence (P=0.207)$(P = 0.207)$ that, on average, plant emission rates are affected by the type of plant eater.

24.29 **(a)** ANOVA is risky. The standard deviation ratio is nearly 3 and the samples are small. **(b)** We test whether the mean biomasses from the three conditions are the same versus at least one population mean is different. **(c)** F=27.52,$F = 27.52$, df=2 and 15, $\text{df} = 2 \text{ and } 15$, P<0.0005$P < 0.0005$ (highly significant), but use with caution as noted in (a).

24.31 H0: H_0: Relative fitness is not influenced by previous pH exposure. Full data set: standard deviations = 0.058, 0.023, and 0.115; F=48.571,$F = 48.571$, df=2 and 15,$\text{df} = 2 \text{ and } 15$, P=2.8×10−7. $P = 2.8 \times 10^{-7}$. Without the low outlier (0.56): standard deviations = 0.058, 0.023, and 0.055; F=93.329, df=2 and 14, $F = 93.329$, $\text{df} = 2 \text{ and } 14$, P=8.1×10−9,$P = 8.1 \times 10^{-9}$, highly significant.

24.33 The differences among the groups were significant (F=3.44,$(F = 3.44$, df=3 and 27,$\text{df} = 3 \text{ and } 27$, P=0.031).$P = 0.031)$. Nitrogen had a positive effect, the phosphorus and control groups were similar, and the plants that got both nutrients fell between the others.

24.35 Means = 21.5, 15.5, 15.7, 15.5, and 15.6; standard deviations = 0.74, 0.52, 0.49, 0.52, and 0.51. ANOVA conditions satisfied. (Note: Time to heal is continuous, but the researchers rounded the values to the nearest day. We treat these values as continuous.) ANOVA: F=332.50,$F = 332.50$, df=4 and 70,$\text{df} = 4 \text{ and } 70$, P≈0,$P \approx 0$, highly significant. The placebo group performed worse than all the other treatment groups.

24.37 x̄=21.585,MSG=745.5,MSE=460.2, df=3 and 28, $\bar{x} = 21.585$, $\text{MSG} = 745.5$, $\text{MSE} = 460.2$, $\text{df} = 3 \text{ and } 28$, F=1.62,$F = 1.62$, P=0.207, $P = 0.207$, not statistically significant.

Chapter 25

25.1 Individuals: elderly patients with PAD. Response variable: maximum walking distance. Two treatments. H0: μD-M=0, Ha: μD-M>0.$H_0\colon \mu_{D-M}=0,\ H_a\colon \mu_{D-M}>0.$ Matched pairs *t* test.

25.3 Individuals: overweight adolescents with type 1 diabetes and treated with insulin. Response variable: body mass index. Two treatments. H0: μM=μP,Ha: μM<μP.$H_0\colon \mu_M=\mu_P, H_a\colon \mu_M<\mu_P.$ Two-sample *t* test.

25.5 Individuals: water bottles. Response variable: BPA concentration. Explanatory variable: storage time. H0: β=0, Ha: β>0. $H_0\colon \beta=0,\ H_a\colon \beta>0.$ *t* test for the regression slope.

25.7 **(a)** ANOVA. H0: μ1=μ2=μ3=μ4,$H_0\colon \mu_1=\mu_2=\mu_3=\mu_4,$ Ha: $H_a\colon$ At least one condition yields a different mean ZIKV quantity. **(b)** Two-sample *t* test. H0: μ25HC=μplacebo,Ha: μ25HC<μplacebo. $H_0\colon \mu_{25HC}=\mu_{placebo}, H_a\colon \mu_{25HC}<\mu_{placebo}.$ **(c)** Two possible approaches. (1) Two-sample *z* test for proportions. H0: p25HC=pplacebo,Ha: p25HC<pplacebo.$H_0\colon p_{25HC}=p_{placebo}, H_a\colon p_{25HC}<p_{placebo}.$ (2) Chi-square test for two-way tables. H0:$H_0\colon$ No relationship between treatment and microcephaly outcome. Ha: H0$H_a\colon H_0$ is not true.

25.9 **(a)** Two-sample *t* test. H0: μS=μU,Ha: μS>μU. $H_0\colon \mu_S=\mu_U, H_a\colon \mu_S>\mu_U.$ **(b)** Matched pairs *t* test. H0: μF-M=0, Ha: μF-M≠0.$H_0\colon \mu_{F-M}=0,\ H_a\colon \mu_{F-M}\neq 0.$

25.11 **(a)** A placebo allows researchers to account for any psychological benefit (or detriment) the subject might get from taking a pill. **(b)** Neither the subjects nor the researchers who worked with them knew who was getting ginkgo extract; this prevents expectations or prejudices from affecting the evaluation of the effectiveness of the treatment. **(c)** t=2.147,$t=2.147,$ for which P=0.0387$P=0.0387$ (df=35.35).$(\mathrm{df}=35.35).$ Those who took gingko extract had significantly more misses per line.

25.13 **(a)** The relationship is clearly linear and very strong; r2r^2 is 0.95. **(b)** 4.270±0.118.$4.270\pm 0.118.$ **(c)** Fish 143 is very wide for its length and has a larger residual than other points; it can make inference less accurate.

25.15 **(a)** The diabetic potentials appear to be large. **(b)** H0: μD=μN, $H_0\colon \mu_D=\mu_N,$ Ha: μD≠μN,$H_a\colon \mu_D\neq\mu_N,$ t=3.077,$t=3.077,$ P=0.0039 (df=36.60).$P = 0.0039\ (\mathrm{df} = 36.60)$. **(c)** t=3.841,$t = 3.841$, P=0.0005 (df=37.15).$P = 0.0005\ (\mathrm{df} = 37.15)$.

25.17 A dotplot or stemplot does not suggest any strong departures from Normality. We find x̄=25.67 and s=8.324.$\bar{x} = 25.67$ and $s = 8.324$. With df=17,$\mathrm{df} = 17$, t*=2.110,$t^* = 2.110$, which gives the interval 25.67±4.14.25.67 ± 4.14. We are 95% confident that the mean healing rate for all newts of this species is between 21.53 and 29.81 micrometers per hour.

25.19 *μ* is the mean difference in healing rates, experimental minus control. H0: μ=0,$H_0\colon \mu = 0$, Ha: μ<0.$H_a\colon \mu < 0$. The dotplot is roughly Normal. t=−2.02,$t = -2.02$, df=13, P=0.032,$\mathrm{df} = 13$, $P = 0.032$, reasonably statistically significant.

25.21 H0: pyellow=3⁄4;ptan=1⁄4, Ha:H0$H_0\colon p_{yellow} = 3/4; p_{tan} = 1/4$, $H_a\colon H_0$ is not true. X2=1.165,$X^2 = 1.165$, df=1,$\mathrm{df} = 1$, P=0.281,$P = 0.281$, not significant. The data are consistent with Mendel’s first law.

25.23 The data are appropriate for ANOVA. ANOVA: F=47.19,$F = 47.19$, df=4 and 85,$\mathrm{df} = 4$ and 85, P<0.0005,$P < 0.0005$, significant. Mean ventilation rate is not the same at all temperatures. A graph of the data suggests that ventilation rate is particularly high at the highest temperature (25 degrees Celsius).

Some Studies Recurring Across Chapters

Animal behavior and physiology

Aphid righting behavior: Examples 19.6, 21.2, 21.6, 21.9, Exercises 19.13, 22.2, 22.4, 22.6, 22.8, 22.10

Chickadee alarm calls: Exercises 3.5, 3.10, 4.2, 4.8

Commercial and transgenic chickens: Examples 11.6, 18.5, Exercise 21.34

Grading oysters: Exercises 3.26, 3.27, 4.27, 23.28, 23.29, 23.30, 23.31

Magellanic penguins: Example 3.6, Exercises 4.5, 4.7

Manatee deaths: Examples 3.3, 3.4, 4.3, Exercises 1.29, 4.10, 6.5, 8.2, 9.27

Monitoring rabies: Exercises 9.7, 12.43

Neural mechanism of nociception: Exercises 3.31, 28.9, 28.10, 28.12

Paw preference in tree shrews: Examples 1.1, 6.4, Exercises 8.3, 15.2

Rats on a cafeteria-style diet: Examples 24.1, 24.2, 24.3, 24.4, 24.6, 26.1, 26.2, 26.3, Exercise 26-6

Sex and longevity in fruit flies: Examples 28.1, 28.2, 28.3, 28.6, 28.10, 28.11, Exercises 11.22, 11.23, 11.24, 18.42, 24.40

Spider silk: Exercises 2.1, 2.5, 2.9, 2.40, 17.2, 17.5, 17.32

Ecosystems and plant physiology

Cicadas as fertilizer: Exercises 2.35, 27.43

Tropical flowers: Exercises 1.39, 25.22

Logging in the rainforest: Exercises 2.37, 11.50, 11.51, 24.5, 24.7, 24.28

Ecological approach to algal bloom control: Exercises 3.21, 3.23, 23.1, 23.3, 23.5, 23.7, 23.9

Mycorrhizal colonies and plant nutrition: Example 26.12, Exercises 26.9, 28.37

Weeds among the corn: Examples 27.1, 27.2, 27.3, 27.4, 27.10, 27.11, Exercise 27.8

Health and medicine

Aging and the blood–brain barrier: Examples 23.1, 23.2, 23.3, 23.5, 23.8, Exercise 8.6

Antibiotic resistance (gonorrhea): Exercises 12.5, 12.6, 12.7, 13.23, 13.24, 13.25

Antibiotic resistance (strep): Exercises 19.1, 19.5, 19.9

Blood types: Examples 9.5, 9.7, 12.1, 12.4, 13.9, 13.10, Exercises 9.16, 9.18, 10.30, 10.32, 12.4, 12.44

Cannabis to treat brain cancer: Exercises 7.2, 7.28, 24.2, 24.9

Children with progeria: Examples 17.3, 17.4, Exercise 27.41

Concussion and sleep: Exercises 1.2, 18.4, 18.6, 25.6

Echinacea for the common cold: Exercises 5.9, 5.10, 5.11, 5.12, 5.13, 5.14, 5.15, 15.25, 18.7, 20.3, 22.37

Essential tremor: Examples 7.8, 7.10, Exercises 8.14, 17.15, 17.36, 25.26

HPV infections: Examples 19.1, 19.3, Exercises 9.15, 22.23

HPV and cervical cancer: Exercises 4.43, 6.14, 20.4, 20.44, 22.30

Influenza: Exercises 6.37, 13.2, 19.11

Men’s and women’s physiology: Example 4.8, Exercise 28.11

Mercury in pregnancy: Exercises 2.4, 2.38, 11.2, 17.17

Myopia in children: Exercises 5.28, 8.8

Nanomedicine: Exercises 1.41, 2.31, 27.46

Peanut allergies: Exercises 5.1, 5.6, 7.7, 20.24, 20.26, 22.25

Pertussis epidemic: Examples 1.3, 1.7, 6.11, 18.1, Exercise 1.9

Smoking cessation (abrupt versus gradual): Examples 19.2, 20.1, 20.2, Exercise 20.1

Testosterone and obesity in adolescent males: Example 17.2, Exercises 18.26, 23.37, 23.39

Human behavior

Bedtime electronic reading: Exercises 7.8, 17.26, 17.27

Children’s food choices: Exercises 1.3, 1.4, 19.39, 20.29

Contact lens use: Example 13.8, Exercise 19.28

Daily activity and obesity: Examples 18.2, 18.3, 18.4, Exercises 2.36, 27.1, 27.3, 27.7, 27.9

Facebook experiment: Examples 15.4, 15.5, Exercise 7.13

Fidgeting (NEA) and body weight: Examples 3.5, 4.1, 4.2, 4.3, Exercises 23.13, 23.14, 23.15, 23.16, 23.17, 23.18, 23.19, 23.20, 23.21, 23.22

Highly Superior Autobiographical Memory: Example 2.11, Exercises 1.32, 1.34, 2.13, 18.5, 18.13, 18.14

Laughing groups: Exercises 2.28, 9.20, 9.21, 9.22, 9.23

Night watch (nighttime brain activity): Exercises 3.6, 3.24, 23.42

Polarizing taste of cilantro: Exercises 9.49, 10.21, 10.22, 10.23, 10.24, 19.38, 19.40, 20.11, 20.12, 20.13, 20.14, 20.15

Tongue piercing and dental health: Exercises 18.31, 22.31

Large data sets

Ant morphology: Exercises 8.18, 16.28, 25.25

Butterfly wings: Exercises 8.15, 17.7, 25.27, 25.28

Children’s growth: Exercises 4.46, 25.30

Elderly health: Exercises 2.42, 17.47, 18.43, 24.39,

Estimating body fat in men: Exercises 8.16, 8.17, 16.27, 17.46, 19.42, 23.43, 23.44

Everglades: Exercises 1.45, 3.40

Lung capacity of children: Exercises 8.19, 8.20, 11.52, 25.29, 25.31

INDEX

A

68–95–99.7 rule, 274

95% confidence, 351

Addition rule for probabilities,

  for disjoint events, 223

  general rule, 248

Alternative hypothesis, 357

  nondirectional, 529, 549, 607

  one-sided or one-tailed, 357, 365

  two-sided or two-tailed, 357, 365

Analysis of variance (ANOVA), 610

  F statistic, 610, 623

  F test, 607, 611

  one-way, 610, Chapter 24

  table, 623

Anecdote, xxii

Applet

  Central Limit Theorem, 333

  Confidence Interval, 351, 352

  Correlation and Regression, 91, 104, 121, 122

  Law of Large Numbers, 340, 341

  Mean and Median, 43, 62

  Normal Approximation to Binomial, 309

  Normal Density Curve, 297

  One-Variable Calculator, 15, 16

  One-Way ANOVA, 611

  Probability, 218, 219, 240

  P-value for a Test of One Proportion, 486

    P-value of a Test of Significance, 360, 361, 363, 387
    Simple Random Sample, 149, 161, 170, 172, 192, 193, 194, 196, 329
    Statistical Power, 393, 401
Association
    and causation, 111
    linear, 78. *See also* Correlation
    negative or positive, 72

B

Bar graph, 9
    stacked, 12
Bayes’s theorem, 258, 261
Bias, 147. *See also* Unbiased estimator
Bimodal, 19
Binomial coefficient, 302
Binomial distribution. *See* Distribution, binomial
Binomial setting, 299
Bivariate, 69
Block, 173
Boxplot, 51. *See also* Five number summary
    modified, 54

C

Case-control study, 156
Categorical data. *See* Variable, categorical
Cell, in a two-way table, 546
Cell counts for chi-square, 557
Center of a distribution, 16, 328. *See also* Mean; Median
Central limit theorem, 333
Chi-square distribution. *See* Distribution, chi-square
Chi-square statistic, 529, 552
    terms of, or components, 529, 552
    Pearson’s, 555

Clusters
    in a distribution, 19
    in a scatterplot, 81
Cohort study, 157
Complement, 223
Conditional distribution. *See* Distribution, conditional
Conditional probability. *See* Probability, conditional
Conditions for inference
    ANOVA, 618, 619
    chi-square for goodness of fit, 536, 537
    chi-square for two-way table, 557
    t test for one mean, 418, 435
    t test for regression, 577, 591
    t test for two means, 446, 460
    z test for one mean (simple conditions), 348, 354, 365, 376
    z test for one proportion, 475, 477, 480, 485
    z test for two proportions, 498, 501, 505
Confidence interval, 351, 354
    four-step process, 354
    plus four for difference of proportions, 498
    plus four for one proportion, 477
    relationship to tests, 367
    t for difference of means, 451
    t for matched pairs, 431
    t for one mean, 422
    t for regression mean response, 590
    t for regression slope, 586
    z for difference of proportions (large sample), 486
    z for one mean, 354
    z for one proportion (large sample), 467
Confidence level, 351
Confounding, 143

Continuity correction, 309

Controlling in experiments, 169, 171

Correlation, 78, 574, Chapter 3

- cautions about, 80
- Pearson's, 78
- squared (r^2), 100
- test for, 583

Count, 336

Critical value, 353

- of chi-square distribution Table D
- of correlation *r*, Table E
- of *F* distribution, Table F
- of standard Normal distribution, Table A and Table C
- of *t* distribution, Table C

Cumulative proportion or probability, 278, 304

Cycle, in a time plot, 23

D

Data analysis, *See also* Exploratory data analysis

Data production, Chapters 6, 7

Degrees of freedom

- of ANOVA *F* statistic, 612
- of chi-square distribution, 530, 531, 555
- of *F* distribution, 612
- of regression *s* and *t*, 579, 581
- of standard deviation *s*, 49
- of *t* distribution, 419
- of pooled two-sample *t*, 462
- of two-sample *t*, 450

Dependent,

- events in probability, 243
- variable, 68

Density curve, 225

Design,
  block, 186
  crossover, 175
  cross-sectional, 152
  experimental, Chapter 7
  longitudinal, 157
  matched pairs, 188, 423
  observational, Chapter 6
  sampling, 144, 146
Deviation,
  from distribution pattern, 16
  from scatterplot pattern, 71
Diagnostic test, 258, 262
  positive predictive value (*PPV*), 259
  sensitivity, 258
  specificity, 258
Direction of a relationship, 71
Discrete,
  probability models, 224
  random variable, 229
  sample space, 221
  variable, 5
Disjoint events, 223, 241
Distribution,
  binomial, 300
    Normal approximation for, 309, 338
    continuity correction, 309
  chi-square, 530, 555, Table D
  conditional, 127, 546
  correlation, Table E
  F, 612 , Table F
  geometric, 322

- marginal, 124, 127
- multinomial, 322
- Normal, 273, Chapter 11
  - standard, 277, 285, Table B
- Poisson, 311
- population, 330
- probability. *See* Probability distribution
- of categorical variable, 8
- of quantitative variable, 8, 16, 17
- of random variable, 229
- of variable, 8, Chapter 1
- sampling. *See* Sampling distribution
- skewed (to the right, to the left), 17
- symmetric, 17
- *t*, 419, Table C
- uniform, 228
- unimodal, 17

Dotplot, 20

E

EESEE, 210, 413, 650

Effect size, 384, 392, 512

Error bars, 51

Estimation, 348. *See also* Confidence interval

Ethics, 182

Event, 220

Expected count, 528, 537, 550, 557

Experiment, 112, 142
- block design, 173
- cautions about, 178
- completely randomized, 173
- crossover, 175
- design of, Chapter 7

  double-blind, 179
  double-dummy, 179
  matched pairs design, 175, 430
  randomized comparative, 170
  repeated measures, 176
  single-blind, 179
Experimental group, 169
Experimental units, 165
Explanatory variable. *See* Variable, explanatory
Exploratory data analysis, 7
Extrapolation, 108

F

F distribution. *See* Distribution, *F*
F statistic 610, 623, Table F
F test for ANOVA, 610, Chapter 24
Factor, experimental, 165
Factorial (*n*!), 303
Five-number summary, 47
Form of a relationship, 71
Four-step process,
  for confidence interval, 354
  for hypothesis test, 365
  for organizing a statistical problem, 58
Frequency, 9
  relative, 9

G

Geometric distribution. *See* Distribution, geometric
Goodness of fit, *See* Hypothesis test

H

Histogram, 13
  interpreting, 16

Hypothesis test, 356, Chapters 14–15
  chi-square for goodness of fit, 531
  chi-square for two-way table, 555
  F test in ANOVA, 607, 623
  for homogeneity, 560
  for independence, 560
  for lack of correlation, 583
  four-step process, 365
  power of, 393, 395
  relationship to confidence intervals, 367
  t test for matched pairs, 431
  t test for one mean, 426, 435
  t test for regression slope, 581
  t test for two means, 451
    pooled, 462, 624
  z test for one mean, 364
  z test for one proportion, 485, 533
  z test for two proportions, 505, 561

I

Independence
  of events in probability, 243, 251
  of random samples, 446, 618
  test of, 560
Individuals, 4, 165
Inference, statistical, 146, 347, Chapters 14–24
  in practice, Chapter 15
Influential observation, 103
Informed consent, 183, 184
Institutional review board, 184, 187
Intercept of a line, 93, 96, 578
Interquartile range (*IQR*), 45

L

Lack of realism, 180

Large numbers, law of, 340

Least-squares regression, 95, 96, 99, 575

Linear relationship, 71, 78. *See also* Correlation

  testing, 581

Logarithm transformation, 88, 105

Lurking variable. *See* Variable, lurking

M

Margin of error, 351, 380, 381

  choosing sample size, 391, 483

Marginal distribution. *See* Distribution, marginal

Matched pairs design, 175, 430

Mean

  of binomial distribution, 306

  of Normal distribution, 271

  of Poisson distribution, 312

  of a population (μ), 324

  response in regression, 577, 590

  of a random variable, 230

  of a sample (x-$\bar{x}$), 41, 324, 334

  of sample mean, 329

  of sample proportion, 338

Mean squares, 622

  for error (*MSE*), 622, 623

  for groups (*MSG*), 622, 623

Median, 40

Meta-analysis, 462

Model, 94

Multiple comparisons, 549, 607

Multiplication rule for probabilities

  independent events, 243

  general rule, 250

Multivariate, 75, 108

N

Nonresponse, 153

Normal distribution. *See* Distribution, Normal

Normal quantile plot, 289

  granularity in, 298

Null hypothesis, 357

Number needed to treat (*NNT*), 512

O

Observational study, 142

  case control, 156

  cohort, 157

  cross-sectional, 152

  longitudinal, 157

  prospective, 156

  retrospective, 156

Odds, 232

  odds ratio, 509

Outlier,

  1.5 ×IQR$1.5 \times IQR$ rule for suspected, 53

  in distribution, 16

  in regression, 71

P

Parameter, 39, 324

Pattern

  of a distribution, 16

  of a scatterplot, 71

Percentile, 41, 282

Personal probability. *See* Probability, personal

Pie chart, 9

Placebo, 169
Plus four
  confidence intervals, 480, 501
  estimate, 480
Pooled estimator
  of population proportion, 505
  of standard deviation in ANOVA, 624
  of variance in two samples, 462
Population, 4, 144
Population regression line, 577
Positive predictive value (*PPV*), *See* Diagnostic test
Power (of *z* test) 393, 395
Prediction
  inference for, 587
  interval, 590
Probability, 215, Chapters 9, 10
  binomial, 303
  conditional, 246
  frequentist, 218
  Normal, Chapter 11
  personal, 218
  Poisson, 312
  rules of, 223, 243, 248, 250
Probability distribution, 229
Probability model, 220
  continuous, 224, 227
  discrete, 224
Probability sample. *See* Sample, probability
Proportion, population (p), 324, 338, 467, 486
Proportion, sample (p^$\hat{p}$), 324, 336, 474
  distribution of, 338
  pooled, 505

P-value, 360, 365, 383, 386

Q

Quartiles (first and third), 45
    interquartile range, *IQR*, 45

R

Random digits, table of, 149, Table A
Random variable, 229
    continuous, 230
    discrete, 229
    expected value (mean μμ) of, 230
    standard deviation (σσ) of, 230
    variance (σ2σ^2) of, 230
Randomization, in experiments, 171
Randomized comparative experiment. *See* Experiment
Randomness, 215
Range, 45
Raw data, 20
Regression, Chapter 4
    cautions about, 107
    inference for, Chapter 23
Regression line, 94
    population, 577
Regression standard error. *See* Standard error, regression
Replication, 179
Residual, 95, 578
    plot, 591
Resistant measure, 43
Response bias, 154
Response variable. *See* Variable, response
Risk, 232
    absolute risk reduction (*ARR*), 512

  - relative risk (*RR*), 509
  - relative risk reduction (*RRR*), 512

Robustness, 418, 435, 618

Roundoff error, 9, 125

S

Sample, 4, 144
  - convenience, 147
  - design of, 144, 146
  - multistage, 150
  - probability, 149
  - simple random (SRS), 149
  - stratified, 150
  - systematic, 150

Sample size, choosing
  - in confidence interval for a mean, 391
  - in confidence interval for a proportion, 483

Sample space, 220

Sample survey, 152

Sampling design. *See* Sample, design of

Sampling distribution, 327, Chapter 13
  - of difference of means, 450
  - of difference of proportions, 496
  - of sample count, 301
  - of sample mean, 330
  - of sample proportion, 338

Scatterplot, 69, 574
  - categorical variables in, 75
  - interpreting, 71

Scientific approach, 388

Sensitivity. *See* Diagnostic test

Shape of a distribution, 16, 327

Significance, statistical, 171, 360, 384, 385

Significance level, 362, 395

Significance test. *See* Hypothesis test

Simple conditions. *See* Conditions for inference

Simple random sample (*SRS*). *See* Sample, simple random

Simpson’s paradox, 132

Simulation, 325

Skewed distribution. *See* Distribution, skewed

Slope of a line, 93, 96, 578

Specificity, *See* Diagnostic test

Spread of a distribution, 16, 328. *See also* Quartiles, Standard deviation

Spreadsheet, 4.

SRS. *See* Sample, simple random

Standard deviation, 47
  condition for ANOVA, 618, 619
  inference for, 462
  of binomial distribution, 306
  of difference of sample means, 450
  of difference of sample proportions, 497
  of Normal distribution, 271
  of Poisson distribution, 312
  of a population (σσ), 333, 618
  of a random variable (σσ), 230
  of sample mean, 329
  of sample proportion, 338
  of population regression line, 577
  pooled, 624

Standard error, 418
  of difference of sample means, 450
  of difference of sample proportions, 497
  of individual regression response, 590
  of least-squares slope, 581, 586
  of mean regression response, 590

  regression, 578
  of sample mean, 418
  of sample proportion, 476
Standard Normal distribution. *See* Distribution, Normal, standard
Standardized value. *See* z-score
Statistic, 39, 324
Statistical thinking, xxii
Stem and leaf plot. *See* Stemplot
Stemplot, 34
Strength of a relationship, 71. *See also* Correlation
Subjects, in experiments, 165
Sums of squares, 623
Symmetric distribution. *See* Distribution, symmetric

T

t distribution. *See* Distribution, *t*
t inference procedures
  for matched pairs, 431
  for one mean, 422, 426, 435, Chapter 17
  for prediction in regression, 590
  for regression slope, 581, 586
  for two means, 451, Chapter 18
  for two means, pooled, 462, 624
t statistic
  one-sample, 419, 426
  regression slope, 586
  pooled two-sample, 462
  two-sample, 450
Test of significance. *See* Hypothesis test
Test statistic, 360, 365
Testing hypotheses. *See* Hypothesis test
Time plot, 23
  cycles in, 23

- trend in, 23

Time series data, 25

Treatment, in experiments, 165

Tree diagram, 255

Trend, in a time plot, 23

Two-sample problems, 445, 495

Two-way table, 124, 545, Chapter 5

Type I and II errors, 383

U

Unbiased estimator, 330

Undercoverage, 152

Unimodal, 17

- single-peaked, 17

V

Variable, 4. *See also* Random variable

- categorical, 5, Chapter 5
- continuous, 5
- dependent, 68
- discrete, 5
- explanatory, 68
- independent, 68
- lurking, 108, 142
- nominal, 5
- ordinal, 5
- quantitative, 5
- response, 68
- row and column, 124

Variance, 47, 230

Venn diagram, 241

Voluntary response sample, 147

W

Wording of questions, 154

X

X2χ^2 (chi-square) statistic, 529, 552

Z

z-score, 276

z statistic
  for one proportion, 485
  one-sample for mean, 364
  for two proportions, 505

SYMBOLS

μμ	population mean
x¯$\bar{x}$	sample mean
σ2σ^2	population variance
s2s^2	sample variance
σσ	population standard deviation
ss	sample standard deviation
Q1,Q3Q_1, Q_3	first and third quartiles in a sample
pp	population proportion
p ^$\hat{p}$	sample proportion
μy =α +βx$\mu_y = \alpha + \beta x$	αα slope and ββ intercept of a population regression line;
	μyμ_y mean response for a given value of xx
y=a+bx$y = a + bx$	aa slope and bb intercept of a sample regression line
rr	correlation coefficient for a sample linear relationship
r2r^2	square of the correlation coefficient
P(A)$P(A)$	probability of event AA
P(A|B)$P(A \mid B)$	probability of event AA, given that event BB has occurred or is true
N(μ,σ)$N(\mu, \sigma)$	Normal distribution with mean μμ and standard deviation σσ
zz-score	standardized score under a Normal distribution
SRS	simple random sample
SEM	standard error of the mean
SE	standard error of an estimate
df	degrees of freedom
mm	margin of error in a confidence interval
H0H_0	null hypothesiss
HaH_a	alternative hypothesis
αα	chosen significance level and probability of a Type I error
ββ	in a hypothesis test, probability of a Type II error
zz	zz statistic
tt	tt statistic

X2X^2	chi-square statistic
FF	FF statistic

CHAPTER 26 More about Analysis of Variance: Follow-up Tests and Two-Way ANOVA

NHPA/SuperStock

IN THIS CHAPTER WE COVER...

- Beyond one-way ANOVA
- Follow-up analysis: Tukey pairwise multiple comparisons
- Follow-up analysis: contrasts
- Two-way ANOVA: conditions, main effects, and interaction
- Inference for two-way ANOVA
- Some details of two-way ANOVA

Analysis of variance (ANOVA) is a statistical method for comparing the means of several populations based on independent random samples, or the mean responses to several treatments in a randomized comparative experiment. When we compare just two means, we use the two-sample *t* procedures described in Chapter 18. ANOVA allows comparison of any number of means.The basic form of ANOVA is one-way ANOVA, which treats the means being compared as mean responses to different levels of a single variable. For example, in Chapter 24 we used one-way ANOVA to compare the mean weights of adult male Wistar rats fed one of three types of diets. Figure 26.1 shows the Minitab ANOVA output for these data (displayed in Table 24.1, page 606).

BEYOND ONE-WAY ANOVA

You should recall or review the big ideas of **one-way ANOVA** from Chapter 24. One-way ANOVA compares the means $\mu_1, \mu_2, \ldots, \mu_k$ of k populations based on samples of sizes $n_1, n_2, \ldots, n_k$ from these populations.

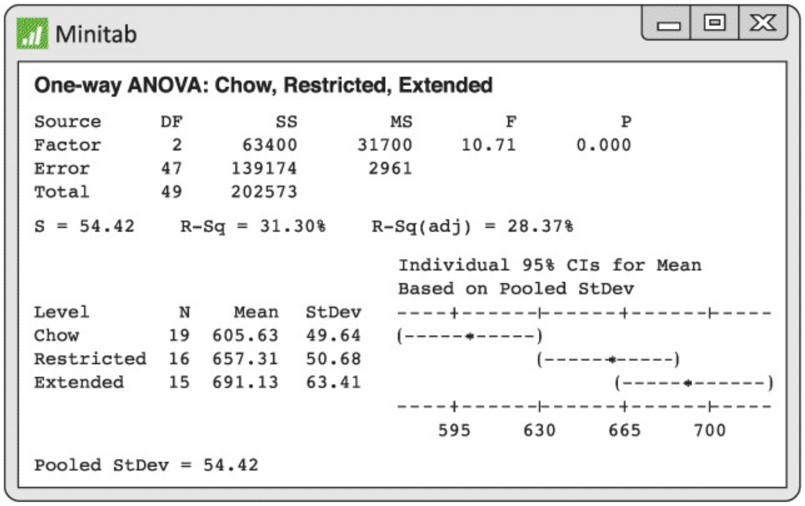
- Using separate two-sample *t* procedures to compare many pairs of means is a bad idea, because we don’t have a *P*-value or a confidence level for the complete set of comparisons together. This is the problem of **multiple comparisons.**
- One-way ANOVA gives a single test for the null hypothesis that all the population means are the same against the alternative hypothesis that not all are the same (H_0 simply is not true).
- ANOVA works by comparing how far apart the sample means are relative to the variation among individual observations in the same sample. The test statistic is the **ANOVA *F* statistic**

$$F = \frac{\text{variation among the sample means}}{\text{variation among individuals in the same sample}}$$

  The *P*-value comes from an ***F* distribution.**
- The required **conditions for ANOVA** are *independent random samples* from each of the *k* populations (or a randomized comparative experiment with *k* treatments), *Normal distributions* for the response variable in each population, and a *common standard deviation* σ in all populations. Fortunately, ANOVA inference is quite robust against moderate violations of the Normality and common standard deviation conditions.
- In basic statistical practice, we combine the *F* test with descriptive data analysis to check the conditions for ANOVA and to see which means appear to differ and by how much.

Examples 24.1, 24.2, and 24.3 (pages 605, 608, and 609) showed all the steps required for a one-way ANOVA. This chapter moves beyond basic one-way ANOVA in two directions.

Follow-up analysis. The ANOVA *F* test in Figure 26.1 tells us only that the population means are not the same. We would like to say which means differ and by how much. For example, do the data allow us to say that the "extended diet" population does have a higher mean weight than the "chow diet" and the "restricted diet" populations of adult male Wistar lab rats? This is a *follow-up analysis* to the *F* test that goes beyond data analysis to confidence intervals and tests of significance for specific comparisons of means.



Minitab

One-way ANOVA: Chow, Restricted, Extended

```
Source      DF        SS        MS        F        P
Factor       2     63400     31700    10.71    0.000
Error       47    139174      2961
Total       49    202573

S = 54.42    R-Sq = 31.30%    R-Sq(adj) = 28.37%

                                   Individual 95% CIs for Mean
                                   Based on Pooled StDev
Level        N    Mean   StDev     ----+------+------+------+----
Chow        19  605.63   49.64     (-----*-----)
Restricted  16  657.31   50.68                 (-----*-----)
Extended    15  691.13   63.41                          (-----*------)
                                   ----+------+------+------+----
                                      595    630    665    700

Pooled StDev = 54.42
```

Figure 26.1

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.1 Minitab ANOVA output for the rat weight data of Example 24.3.

Two-way ANOVA. One-way ANOVA compares mean responses for several levels of just one explanatory variable. In Example 24.1, that variable is "the type of diet provided." Suppose that we have data on *two* explanatory variables, say, the type of diet provided and whether the lab rats are physically active. There are now 6 groups formed by combinations of diet type and physical activity, as follows:

		Variable 2	
		Active	**Inactive**
	Chow	Group 1	Group 2
Variable 1	**Restricted**	Group 3	Group 4

	Extended	Group 5	Group 6

One-way ANOVA will still tell us if there is evidence that mean body weight in these 6 experimental groups differs. But we want more: Does diet type matter? Does physical activity matter? And do these two variables *interact?* That is, does the effect of diet type change when the lab rats are physically active? Perhaps physical activity reduces the craving for cafeteria food, so that diet type has less effect when the rats are active than when they are inactive. To answer these questions we must extend ANOVA to take into account the fact that the 6 groups are formed from two explanatory variables. This is *two-way ANOVA.*

We will discuss follow-up analysis in ANOVA first, and then two-way ANOVA. Fortunately, the distinction between one-way and two-way doesn’t affect the follow-up methods we will present. So once you have mastered these methods in the one-way setting, you can apply them immediately to two-way problems.

FOLLOW-UP ANALYSIS: TUKEY PAIRWISE MULTIPLE COMPARISONS

In Example 24.3 we saw that there is good evidence that the mean body weight of adult male Wistar rats is not the same when they are assigned to a diet consisting of chow only, chow plus restricted access to cafeteria food, and chow plus extended access to cafeteria food.[1] The sample means in Figure 26.1 suggest that (as we might expect) the mean body weight is highest in rats given extended access to cafeteria food and lowest in rats given chow only.

EXAMPLE 26.1 Comparing groups: individual *t* procedures

Let’s use A, B, and C to refer to the chow, restricted, and extended groups, respectively. How much higher is the mean body weight of rats given restricted access to cafeteria food than that of rats given chow only? A 95% confidence interval comparing Groups A and B answers this question. Because the conditions for ANOVA require that the population standard deviation be the same in all three populations of rats, we can use a version of the two-sample *t* confidence interval that also assumes equal standard deviations.

Cristian Ciobanu/Alamy

The Minitab output in Figure 26.1 gives the pooled standard deviation (first defined in chapters 18 and 24, pages 462 and 624) as sp=54.42 grams (g).$s_p = 54.42 \text{ grams (g)}$. This is an estimate of the common standard deviation σσ based on all three samples. It has 47 degrees of freedom, the degrees of freedom for “Error” in the ANOVA table. The standard error for the difference in sample means x-A-x-B$\bar{x}_A - \bar{x}_B$ is (page 462) sp1nA+1nB

$$s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$$

A 95% confidence interval for μA-μB$\mu_A - \mu_B$ would therefore be

(x-A-x-B)±t*sp1nA+1nB

$$(\bar{x}_A - \bar{x}_B) \pm t^* s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}$$

using t*=2.012$t^* = 2.012$ from technology (or approximately t*=2.021 $t^* = 2.021$ from Table C for df =40,$\text{df} = 40,$ conservatively, since there is no row for df = 47$\text{df} = 47$ in Table C).

However, we really want to estimate all three **pairwise differences** among the population means,

μA-μB μA-μC μB-μC

$$\mu_A - \mu_B \qquad \mu_A - \mu_C \qquad \mu_B - \mu_C$$

CAUTION

Three 95% confidence intervals will not give us 95% confidence that all three *simultaneously capture their true parameter values.* This is the familiar problem of multiple comparisons that we discussed in Chapters 22 and 24.

In general, we want to give confidence intervals for all pairwise differences among the population means μ1,μ2,…,μk$\mu_1, \mu_2, \ldots, \mu_k$ of *k* populations. We want an **overall confidence level** of (say) 95%. That is, in very many uses of the method, *all* the intervals will simultaneously capture the true differences 95% of the time. To do this, take the number of comparisons into account by replacing the *t* critical value t*t^* in Example 26.1 with another critical value based on the distribution of the difference between the largest and smallest of a set of *k* sample means. We will call this critical value m*,$m^*,$ for multiple comparisons. Values of m*m^* no longer come from a *t* table. They depend on the number of populations we are comparing and on the total number of observations in the samples, as well as on the confidence level we want. Software is very helpful for practical use. This method is named after its inventor, John Tukey (1915–2000), the same man who developed the ideas of modern data analysis. A short table of m*m^* values for a 95% confidence level (Table G) is provided for convenience at the end of this chapter.

TUKEY PAIRWISE MULTIPLE COMPARISONS

In the ANOVA setting, we have independent SRSs of size nin_i from each of *k* populations having Normal distributions with means μiμ_i and a common standard deviation σ.σ. **Tukey simultaneous confidence intervals** for all pairwise differences μi-μj$\mu_i - \mu_j$ among the population means have the form

(x-i-x-j)±m*sp1ni+1nj

$$(\bar{x}_i - \bar{x}_j) \pm m^* s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Here x-i$\bar{x}_i$ is the sample mean of the iith sample and sps_p is the pooled estimate of σ.σ. The critical value m*m^* depends on the confidence level C,C, the number of populations *k*, and the total number of observations N.N.

If all samples are the same size, the Tukey simultaneous confidence intervals provide an overall level *C* of confidence that *all* the intervals simultaneously capture the true pairwise differences. If the samples differ in size, the true confidence level is at least as large as *C*. That is, the conclusions are then conservative.

To carry out **simultaneous tests** of the hypotheses

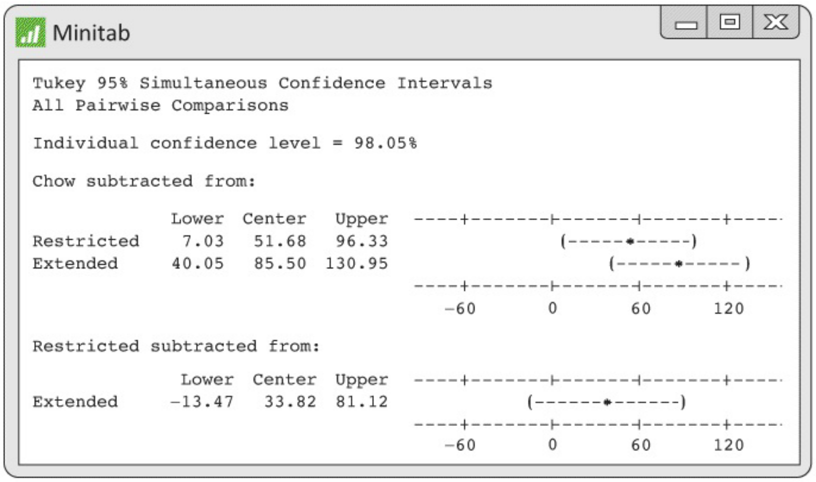
H0: μi =μjHa: μi≠μj

$$\begin{aligned} H_0 &: \mu_i = \mu_j \\ H_a &: \mu_i \neq \mu_j \end{aligned}$$

for all pairs of population means, reject H_0 for any pair whose confidence interval does not contain 0. These tests have overall significance level no less than 1 − *C* That is, 1 − *C* is the probability that, when all the population means are equal, *any* of the tests incorrectly rejects its null hypothesis.

EXAMPLE 26.2 Rats and a cafeteria-style diet: multiple intervals

Figure 26.2 contains more Minitab output for the ANOVA comparing the mean body weights in 3 experimental groups of lab rats. We asked for Tukey multiple comparisons with an overall error rate of 5%. That is, the overall confidence level for the three intervals together is 95%.



```
Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons

Individual confidence level = 98.05%

Chow subtracted from:

             Lower  Center   Upper   ----+--------+-------+-------+----
Restricted    7.03   51.68   96.33                 (-----*-----)
Extended     40.05   85.50  130.95                     (-----*-----)
                                     ----+--------+-------+-------+----
                                        -60       0      60      120

Restricted subtracted from:

              Lower  Center  Upper   ----+--------+-------+-------+----
Extended     -13.47   33.82  81.12              (------*------)
                                     ----+--------+-------+-------+----
                                        -60       0      60      120
```

Figure 26.2

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.2 Additional Minitab ANOVA output showing Tukey pairwise multiple comparisons for the rat weight data, for Example 26.2.

The format of the Minitab output takes some study. Be sure you can see that the Tukey confidence intervals are

7.03to96.33 for μB-μA 40.05to130.95 for μC-μA -13.47to81.12 for μC-μB

$$\begin{array}{rcrcl} 7.03 & \text{to} & 96.33 & \text{for} & \mu_B - \mu_A \\ 40.05 & \text{to} & 130.95 & \text{for} & \mu_C - \mu_A \\ -13.47 & \text{to} & 81.12 & \text{for} & \mu_C - \mu_B \end{array}$$

If you do not have access to technology, these intervals can easily be computed by hand. Let’s see how to obtain the first interval, for μA-μB. $\mu_A - \mu_B$. Table G at the end of this chapter gives values of m*m^* when using an overall confidence level of 95% and various combinations of N N and k.k. Start by finding the right combination of kk comparisons (top row) and N-k$N-k$ degrees of freedom (left margin). In our example, k=3 $k=3$ and N-k=47,$N-k=47,$ so m*=2.434,$m^*=2.434,$ approximately (based on a conservative df=40,$\text{df}=40,$ since df=47$\text{df}=47$ is not available). The interval for μB-μA$\mu_B - \mu_A$ is therefore

(x-B-x-A)±m*sp1nB+1nA=(657.31-605.63)±(2.434)

(54.42)116+119=51.68±44.94=6.74 to 96.62

$$\begin{aligned}(\bar{x}_B - \bar{x}_A) \pm m^* s_p \sqrt{\frac{1}{n_B} + \frac{1}{n_A}} &= (657.31 - 605.63) \pm (2.434)(54.42)\sqrt{\frac{1}{16} + \frac{1}{19}} \\ &= 51.68 \pm 44.94 \\ &= 6.74 \text{ to } 96.62\end{aligned}$$

Notice that the value of m*m^* we use here is larger than the value of t*t^* in Example 26.1. This is the price we pay for having 95% confidence not just in one interval but in all three simultaneously.

EXAMPLE 26.3 Rats and a cafeteria-style diet: multiple tests

The ANOVA null hypothesis is that all population means are equal,

H0:μA=μB=μC

$$H_0 : \mu_A = \mu_B = \mu_C$$

We know from the output in Figure 26.1 that the ANOVA *F* test rejects this hypothesis (F=10.71,$(F = 10.71,$ P<0.0005).$P < 0.0005).$ So we have good evidence that *some* pairs of means are not the same. Which pairs? Look at the simultaneous 95% confidence intervals in Example 26.2. Which of these intervals do not contain 0? If an interval does not contain 0, we reject the hypothesis that this pair of population means are equal.

The conclusions are

We can rejectH0:μB=μAWe can rejectH0:μC=μAWe cannot rejectH0:μC=μB

$$\begin{array}{ll}\text{We } can\ reject & H_0 : \mu_B = \mu_A \\ \text{We } can\ reject & H_0 : \mu_C = \mu_A \\ \text{We } cannot\ reject & H_0 : \mu_C = \mu_B\end{array}$$

That is,

We do have enough evidence to conclude that μA≠μBWe do have enough evidence to conclude that μA≠μCWe do not have enough evidence to conclude that μB≠μC

$$\begin{array}{l}\text{We } \textit{do} \text{ have enough evidence to conclude that } \mu_A \neq \mu_B \\ \text{We } \textit{do} \text{ have enough evidence to conclude that } \mu_A \neq \mu_C \\ \text{We } \textit{do not} \text{ have enough evidence to conclude that } \mu_B \neq \mu_C\end{array}$$

This Tukey simultaneous test of three null hypotheses has the property that when all three hypotheses are true, there is only a 5% probability

> that *any* of the three tests wrongly rejects its hypothesis.

Recall what a test at a fixed significance level such as 5% tells us: either we *do have enough evidence* to reject the null hypothesis, or the data *do not give enough evidence* to allow rejection.

The study found evidence that rats on a chow-only diet differ significantly in body weight from rats given restricted access and from rats given extended access to a cafeteria-style diet. However, the study did not find evidence that restricted and extended access to cafeteria food result in rats with significantly different body weights. That is, x-A=605.63 $\bar{x}_A = 605.63$ and x-B=657.31$\bar{x}_B = 657.31$ are farenough apart to conclude that the population means µAμ_A and µBμ_B differ, x-A=605.63$\bar{x}_A = 605.63$ and x-C=691.13$\bar{x}_C = 691.13$ are farenough apart to conclude that the population means µAμ_A and µCμ_C differ, but x-B=657.31$\bar{x}_B = 657.31$ is not far enoughfrom x-C=691.13$\bar{x}_C = 691.13$ to ruleout the possibility that the population means µBμ_B and µCμ_C might be the same.

Notice that the Tukey method does not give a *P*-value for the three tests taken together. Rather, we have a set of “reject” or “fail to reject” conclusions with an overall significance level that we fixed in advance, 5% in this example. There are several other multiple-comparisons procedures that produce simultaneous confidence intervals with an overall confidence level or simultaneous tests with an overall probability of any false rejection. The Tukey procedures are arguably the most useful.[2] If you can interpret results from Tukey, you can understand output from other multiple-comparisons procedures.

APPLY YOUR KNOWLEDGE

26.1 Caffeine and sugar. Exercise 24.22 (page 628) describes a double-blind randomized experiment that assigned healthy undergraduate students to drink one of four beverages after fasting overnight: water, water with 75 mg of caffeine, water with 75 g of glucose, and water with 75 mg of caffeine and 75 g of glucose. The subjects performed a cognitive task, and their reaction times in the task are summarized below (SEM is the standard error of the mean):[3]

Beverage	*n*	x-$\bar{x}$	SEM
Water	18	389.35	18.50
Water and caffeine	18	320.16	17.98
Water and glucose	18	318.16	17.04
Water, caffeine, and glucose	18	336.44	14.02

An ANOVA *F* test gives a significant *P*-value of 0.0134, with MSE=5186.0358.$\text{MSE} = 5186.0358.$

a. Because all four groups are the same size, the margin of error is the same for all 6 pairwise comparisons. Obtain this margin of error using Table G on page 26-40. Find the Tukey simultaneous 95% confidence intervals for all pairwise comparisons of population means.

b. Explain in simple language what “95% confidence” means for these intervals.

c. Which pairs of means differ significantly at the overall 5% significance level?

26.2 Logging in the rain forest. Exercise 24.5 (page 615) describes a study comparing forest plots in Borneo that had never been logged (Group 1) with similar plots nearby that had been logged 1 year earlier (Group 2) and 8 years earlier (Group 3). The three groups can be considered to be independent random samples. The data appear in Table 24.2 (page 616); the variable Trees is the number of trees in a plot.[4] The one-way ANOVA shown in Figure 24.4 compared the mean counts of trees in the 3 types of forest plots and was statistically significant, with P=0.0002.$P = 0.0002.$ It also gave MSE=27.3574.$\text{MSE} = 27.3574.$

a. Find the Tukey simultaneous 95% confidence intervals for all pairwise comparisons of population means. Use software or Table G on page 26-40.
b. Explain in simple language what “95% confidence” means for these intervals.
c. Which pairs of means differ significantly at the overall 5% significance level?

26.3 Which color attracts beetles best? To detect the presence of harmful insects in farm fields, we can put up boards covered with a sticky material and examine the insects trapped on the boards. Which colors attract insects best? Experimenters placed six boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped. Here are the data:[5]

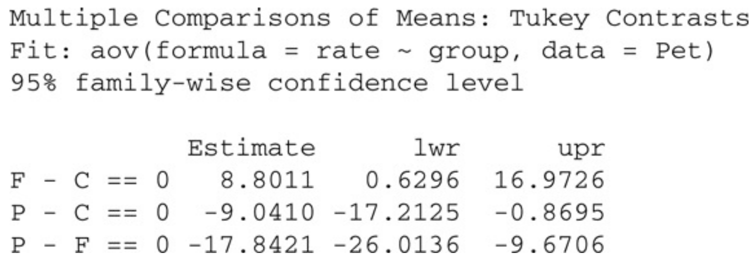
Board color	**Beetles trapped**					
Blue	16	11	20	21	14	7
Green	37	32	20	29	37	32
White	21	12	14	17	13	20
Yellow	45	59	48	46	38	47

ANOVA gives very strong evidence (P<0.0005,$(P < 0.0005,$ MSE = 32.167)$\text{MSE} = 32.167)$ that the colors differ in their ability to attract beetles.

a. How many pairwise comparisons are there when we compare four colors?
b. Use software or Table G on page 26-40 to obtain the Tukey simultaneous 95% confidence intervals for all pairwise comparisons of population means. Which pairs of colors are significantly different when we require a significance level of 5% for all comparisons as a group?

26.4 Dogs, friends, and stress. If you are a dog lover, perhaps having your dog along reduces the effect of stress. The EESEE story “Stress among Pets and Friends” describes a study designed to test this premise. The researchers recruited 45 women who said they were dog lovers, and randomly assigned them to three groups: alone (the control group, C), in the company of a good friend (F), or in the company of their dog (D). Each subject’s mean heart rate while performing a stressful task was recorded. The study found very strong evidence (ANOVA: $F=14.8,P=0.00001$) (ANOVA: $F = 14.8, P = 0.00001$) that mean heart rates under stress do differ depending on whether a pet, a friend, or no one is present.

a. We want to know whether the means for the two treatments (pet, friend) differ significantly from each other and from the mean for the control group. What are the corresponding three null hypotheses?
b. We want to be 95% confident that we don’t wrongly reject any of the three null hypotheses. What can we conclude from the following Tukey pairwise comparisons provided by the software R?



```
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = rate ~ group, data = Pet)
95% family-wise confidence level

           Estimate      lwr      upr
F - C == 0   8.8011   0.6296  16.9726
P - C == 0  -9.0410 -17.2125  -0.8695
P - F == 0 -17.8421 -26.0136  -9.6706
```

FOLLOW-UP ANALYSIS: CONTRASTS[ii]

Multiple-comparisons methods give conclusions about *all* comparisons in some class with a measure of confidence that applies to all the comparisons taken together. For example, Tukey’s method gives conclusions about all pairwise comparisons among a set of population means. These methods are most useful when we did not have any specific comparison in mind before we produced the data.

Multiple-comparisons procedures sometimes give tests or confidence intervals for comparisons that don’t interest us. And they may leave out comparisons that do interest us. If we have *specific questions in mind before we produce data*, it is more efficient to plan an analysis that asks these specific questions.

EXAMPLE 26.4 Which color attracts beetles best?

What color should we use on sticky boards placed in a field of oats to attract cereal leaf beetles? Exercise 26.3 gives data from an experiment in which 24 boards (6 each of blue, green, white, and yellow) were placed at random locations in a field. ANOVA shows that there are significant differences among the mean numbers of beetles trapped by these colors. In Exercise 26.3 you followed ANOVA with Tukey pairwise comparisons.

Holt Studios International/Alamy

But in fact we have specific questions in mind: We suspect that warm colors are generally more attractive than cold colors. That is, *before any data are gathered*, we suspect that blue and white boards will have similar properties, that green and yellow boards will give similar results, and also that the average beetle count for green and yellow will be greater than the average count for blue and white. Therefore, we want to test three hypotheses:

Hypothesis 1	Hypothesis 2	Hypothesis 3
H0:μB=μW $H_0 : \mu_B = \mu_W$	H0:μG=μY $H_0 : \mu_G = \mu_Y$	H0:(μY+μG)/2=(μB+μW)/2 $H_0 : (\mu_Y + \mu_G)/2 = (\mu_B + \mu_W)/2$
Ha:μB≠μW $H_a : \mu_B \neq \mu_W$	Ha:μG≠μY $H_a : \mu_G \neq \mu_Y$	Ha:(μY+μG)/2>(μB+μW)/2 $H_a : (\mu_Y + \mu_G)/2 > (\mu_B + \mu_W)/2$

Two of these hypotheses involve pairwise comparisons. The third does not and also has a one-sided alternative.

We can ask questions about population means by specifying *contrasts* among the means.

CONTRASTS

In the ANOVA setting comparing the means μ1,μ2,…,μk $\mu_1, \mu_2, \ldots, \mu_k$ of *k* populations, a **population contrast** is a combination of the means

L=c1μ1+c2μ2+⋯+ckμk

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$$

with numerical coefficients that add to 0, c1+c2+⋯+ck=0. $c_1 + c_2 + \cdots + c_k = 0.$

EXAMPLE 26.5 Attracting beetles: contrasts

We can restate the three hypotheses in Example 26.4 in terms of three contrasts:

L1=(1)(μB)+(0)(μG)+(-1)(μW)+(0)(μY)L2=(0)(μB)+(1)(μG)+(0)(μW)+(-1)(μY)L3=(-1/2)(μB)+(1/2)(μG)+(-1/2)(μW)+(1/2)(μY)

$$\begin{aligned} L_1 &= (1)(\mu_B) + (0)(\mu_G) + (-1)(\mu_W) + (0)(\mu_Y) \\ L_2 &= (0)(\mu_B) + (1)(\mu_G) + (0)(\mu_W) + (-1)(\mu_Y) \\ L_3 &= (-1/2)(\mu_B) + (1/2)(\mu_G) + (-1/2)(\mu_W) + (1/2)(\mu_Y) \end{aligned}$$

Check that the four coefficients in each line do add to 0. In terms of

> these contrasts, the hypotheses become
>
> | Hypothesis 1 | Hypothesis 2 | Hypothesis 3 |
> |---|---|---|
> | H0:L1=0$H_0 : L_1 = 0$ | H0:L2=0$H_0 : L_2 = 0$ | H0:L3=0$H_0 : L_3 = 0$ |
> | Ha:L1≠0$H_a : L_1 \neq 0$ | Ha:L2≠0$H_a : L_2 \neq 0$ | Ha:L3>0$H_a : L_3 > 0$ |

Some statistical software will test hypotheses and give confidence intervals for any contrasts you specify. Because other software lacks this capability, here’s how to proceed by hand, using information from basic ANOVA output.

To estimate a population contrast

L=c1μ1+c2μ2+⋯+ckμk

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_k\mu_k$$

use the corresponding **sample contrast**

L^=c1x-1+c2x-2+⋯+ckx-k

$$\hat{L} = c_1\bar{x}_1 + c_2\bar{x}_2 + \cdots + c_k\bar{x}_k$$

The sample contrast L^$\hat{L}$ has standard error (estimated standard deviation)

SEL^=spc12n1+c22n2+⋯+ck2nk

$$\text{SE}_{\hat{L}} = s_p\sqrt{\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_k^2}{n_k}}$$

INFERENCE ABOUT A POPULATION CONTRAST

In the ANOVA setting, a level *C* **confidence interval** for a population contrast is

L^±t*SEL^

$$\hat{L} \pm t^*\text{SE}_{\hat{L}}$$

where L^$\hat{L}$ is the corresponding sample contrast and t*t^* is a critical value from the *t* distribution with the degrees of freedom for error in the ANOVA.

To test H0:L=0,$H_0 : L = 0,$ use the *t* statistic

t=L^SEL^

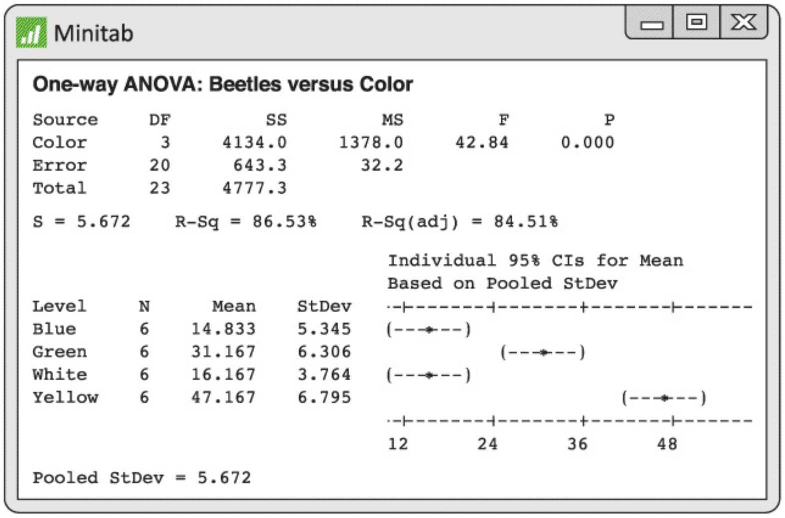
$$t = \frac{\widehat{L}}{\text{SE}_{\widehat{L}}}$$

with the same degrees of freedom.

For one-way ANOVA, the degrees of freedom for error are N-k,$N - k$, where NN is the total number of observations and *k* is the number of populations compared (see Chapter 24, page 612). The box states the result more generally so that it applies to two-way ANOVA as well as one-way. If the contrast is a pairwise difference between means, the contrast confidence interval is exactly the individual confidence interval illustrated in Example 26.1.

EXAMPLE 26.6 Attracting beetles: inference for contrasts

Figure 26.3 displays the Minitab ANOVA output for the study on attracting cereal leaf beetles. The pooled estimate of σσ is sp=5.672, $s_p = 5.672$, and the degrees of freedom for error are 20. Minitab does not offer contrasts, so we must use a calculator.



```
One-way ANOVA: Beetles versus Color

Source   DF      SS      MS      F      P
Color     3  4134.0  1378.0  42.84  0.000
Error    20   643.3    32.2
Total    23  4777.3

S = 5.672   R-Sq = 86.53%   R-Sq(adj) = 84.51%

                               Individual 95% CIs for Mean
                               Based on Pooled StDev
Level   N    Mean  StDev  --+-------+-------+-------+------
Blue    6  14.833  5.345  (---*---)
Green   6  31.167  6.306             (---*---)
White   6  16.167  3.764  (---*---)
Yellow  6  47.167  6.795                         (---*---)
                          --+-------+-------+-------+------
                           12      24      36      48

Pooled StDev = 5.672
```

Figure 26.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.3 Minitab ANOVA output for the study on attracting cereal leaf beetles, for Example 26.6.

The three sample contrasts and their standard errors are

Sample contrast	**Standard error**
L^1=-1.334$\hat{L}_1 = -1.334$	SE1=3.2747$\text{SE}_1 = 3.2747$
L^2=-16.000$\hat{L}_2 = -16.000$	SE2=3.2747$\text{SE}_2 = 3.2747$
L^3= 23.667$\hat{L}_3 = 23.667$	SE3=2.3153$\text{SE}_3 = 2.3153$

Here are the details for the third contrast:

L3=(-1/2)(μB)+(1/2)(μG)+(-1/2)(μW)+(1/2)(μY)L^3=(-1/2)(x̄B)+(1/2)(x̄G)+(-1/2)(x̄W)+(1/2)(x̄Y)L^3=(-1/2)(14.833)+(1/2)(31.167)+(-1/2)(16.167)+(1/2)(47.167)=23.667SE3=(5.672)(-1/2)26+(1/2)26+(-1/2)26+(1/2)26 =(5.672)(0.4082) =2.3153

$$\begin{aligned} L_3 &= (-1/2)(\mu_B) + (1/2)(\mu_G) + (-1/2)(\mu_W) + (1/2)(\mu_Y) \\ \hat{L}_3 &= (-1/2)(\bar{x}_B) + (1/2)(\bar{x}_G) + (-1/2)(\bar{x}_W) + (1/2)(\bar{x}_Y) \\ \hat{L}_3 &= (-1/2)(14.833) + (1/2)(31.167) + (-1/2)(16.167) + (1/2)(47.167) \\ &= 23.667 \\ \text{SE}_3 &= (5.672)\sqrt{\frac{(-1/2)^2}{6} + \frac{(1/2)^2}{6} + \frac{(-1/2)^2}{6} + \frac{(1/2)^2}{6}} \\ &= (5.672)(0.4082) \\ &= 2.3153 \end{aligned}$$

If you use software, your answers may differ slightly due to roundoff error in the hand calculations. A 95% confidence interval for L3L_3 uses t*=2.086$t^* = 2.086$ from [Table C](#) with df=20:$df = 20:$

L^3±t*SE3=23.667±(2.086)(2.3153)=23.667±4.830=18.837 to 28.497

$$\begin{aligned} \hat{L}_3 \pm t^* \text{SE}_3 &= 23.667 \pm (2.086)(2.3153) \\ &= 23.667 \pm 4.830 \\ &= 18.837 \text{ to } 28.497 \end{aligned}$$

We are 95% confident that the average number of beetles attracted by green and yellow boards exceeds the average for blue and white boards by between about 18.8 and 28.5 beetles per board.

There is very strong evidence that the population contrast L3L_3 is greater than 0.The *t* statistic is

t=L^3SE3=23.6672.3153=10.22

$$t = \frac{\hat{L}_3}{\text{SE}_3} = \frac{23.667}{2.3153} = 10.22$$

with 20 degrees of freedom, and P<0.0005$P < 0.0005$ from Table C. The

> other *t* tests conclude that μBμ_B and μWμ_W do not differ significantly, but that there is a significant difference between μGμ_G and μY.μ_Y.

Our confidence intervals and tests for contrasts are individual procedures for each contrast. If we do inference about three contrasts, such as those in Examples 26.5 and 26.6, we face the problem of multiple comparisons again. That is, we do not have an overall confidence level for all three intervals together. There are more advanced multiple-comparisons methods that apply to contrasts just as Tukey’s method applies to pairwise comparisons.

Do not use contrasts to blindly test any possible hypothesis combination. Contrasts are valid only if you have sound biological reasons to state particular hypotheses before you even look at the data.

CAUTION

APPLY YOUR KNOWLEDGE

26.5 Green versus yellow. Using the Minitab output in Figure 26.3, verify the values for the sample contrast L^2$\hat{L}_2$ and its standard error given in Example 26.6. Give a 95% confidence interval for the population contrast L2.L_2. Carry out a test of the hypothesis H0: L2=0$H_0 : L_2 = 0$ against the two-sided alternative. Be sure to state your conclusions in the setting of the study.

26.6 Rats and a cafeteria-style diet: contrasts. Figure 26.1 gives basic ANOVA output for the study of the effects of diet type on rat body weights described in Example 26.1. We might describe the overall effect of access to cafeteria food by comparing the mean body weight for rats given chow only (Group A) with the average of the mean body weights for the two groups of rats with access (restricted or extended) to cafeteria food (Groups B and C).

a. What population contrast *L* expresses this comparison?
b. Starting from the output in Figure 26.1, give the sample contrast that estimates *L* and its standard error.

c. Is there good evidence that the mean weight of rats given chow only is lower than the average for the two groups of rats allowed access to cafeteria food? State the hypotheses in terms of the population contrast L and carry out a test.

d. How much smaller is the mean weight of rats given chow only than the average for the two groups of rats allowed access to cafeteria food? Give a 95% confidence interval.

TWO-WAY ANOVA: CONDITIONS, MAIN EFFECTS, AND INTERACTION

One-way analysis of variance compares the mean responses from any set of populations or experimental treatments when the responses satisfy the ANOVA conditions. Often, however, a sample or experiment has some design structure that leads to more specific questions than those answered by the one-way ANOVA *F* test or by Tukey pairwise comparisons. It is common, for example, to compare treatments that are combinations of values of two explanatory (independent) variables, two **factors** in the language of experimental design. Here is an example.

EXAMPLE 26.7 A two-way layout

The final stages of clinical trials for a new drug often involve finding the most effective dosage and delivery method. The answer may depend on a combination of the two. To investigate this question, randomly assign informed, consenting patients (the *subjects*) to receive the new drug in one of several forms. The drug will be taken either orally or intravenously. Some subjects will receive a low dose, others a medium dose, and others a high dose.

This experiment has two *factors:* delivery method, with 2 values; and dosage, with 3 values. The 6 combinations of one value of each factor form 6 *treatments:*

		Variable *C*		
		Dosage		
		Low	**Medium**	**High**
Variable *R*	**Oral**	Group 1	Group 2	Group 3
Delivery	**Intravenous**	Group 4	Group 5	Group 6

This is a *two-way layout,* with values of one factor forming rows and values of the other forming columns. After treatment completion, each subject is evaluated and his or her condition rated—for example, with a score between 0 and 100 (100 for full recovery). This is the *response*

> *variable.*

To analyze data from such a study, we impose some additional conditions. Here are the **conditions for two-way ANOVA** that will govern our work in this chapter:

1. We have *responses for all combinations* of values of the two factors (all 6 cells in Example 26.7). No combinations are missing in our data. In general, call the two explanatory variables R and C (for Row and Column). Variable R has r different values, and variable C has c different values. The study compares all rc combinations of these values. Such designs are called **crossed,** or fully factorial.
2. In an observational study, we have an *independent SRS* from each of the rc populations. If the study is an experiment, the available subjects are allocated at random among all rc treatments in a *completely randomized design*. In an intermediate design, subjects can represent *independent SRSs* from each of r populations and be separately assigned at random to c treatments in a *randomized block design*. We first saw these different designs in Chapter 8.
3. The response variable has a *Normal distribution* in each population. The population mean responses may differ, but all rc populations have a *common standard deviation* σ.σ.
4. We have the *same number of individuals* n in each of the rc treatment groups or samples. Such designs are called **balanced.**

The second and third conditions are just the usual conditions for ANOVA applied to the two-way layout. Study designs that satisfy the first and second conditions are very common. When you design a study, you can arrange to satisfy the fourth condition; that is, you can choose to have equal numbers of individuals for each treatment. Balanced designs have several advantages in any ANOVA: F tests are most robust against violation of the "common standard deviation" condition when the subject counts are equal or close to equal, and Tukey's method then gives exact overall confidence or significance levels. In the two-way layout, there is an even stronger reason to prefer balanced designs. If the numbers of individuals differ among treatments (an unbalanced design), several alternative analyses of the data are possible. These analyses answer different sets of questions, and you must decide which questions you want to answer. All the sets of questions and all the analyses collapse to just one in the balanced case. This makes interpreting your data much simpler.

To understand the questions that two-way ANOVA answers, return to the drug trial in Example 26.7. In this section we will assume that we know the actual population mean responses for each treatment. That is, we deal with an ideal situation in which we don't have to worry about random variation in the mean responses.

EXAMPLE 26.8 First scenario: main effects with no interaction

Here is the table from Example 26.7 populated with made-up population mean responses to the 6 treatments. (It is unrealistic to expect to know the population means, but our objective in this section is simply to describe possible scenarios.)

		Variable *C* Dosage		
		Low	**Medium**	**High**
Variable *R*	**Oral**	30	45	50
Delivery	**Intravenous**	40	55	60

The mean patient condition scores increase with higher drug dosages and are also higher when the drug is delivered intravenously. The means increase *by the same amount* (10 points) when we move from oral to intravenous drug delivery, *no matter what drug dosage is given.* Turning to the other variable, the effect of moving from low to medium dosage is the same (15 points) for both oral and intravenous delivery, and the effect of moving from medium to high dosage is also the same (5 points) for both delivery methods. Because the result of changing the value of one variable is the same for all values of the other variable, we say that there is *no interaction* between the two variables.

Now average the mean responses for oral and intravenous delivery. The average for oral delivery is

$$30+45+503=1253=41.7$$

$$\frac{30+45+50}{3}=\frac{125}{3}=41.7$$

and the average for intravenous delivery is

40+55+603=1553=51.7

$$\frac{40+55+60}{3}=\frac{155}{3}=51.7$$

Because the averages for the two delivery methods differ, we say that there is a *main effect* for delivery method. Similarly, average the mean responses for each drug dosage. They are 35 for the low dosage, 50 for the medium dosage, and 55 for the high dosage. So changing the drug dosage has an “on average” effect on the response (patient condition score). There is a *main effect* for drug dosage.

Figure 26.4 plots the cell means from Example 26.8. The two solid lines joining low, medium, and high dosages for oral and for intravenous delivery are *parallel.* This reflects the fact that the mean response always increases by 10 points when we move from oral to intravenous delivery, no matter what drug dosage is given. *Parallel lines in a plot of means show that there is no interaction.* It doesn’t matter which variable you choose to place on the horizontal axis.

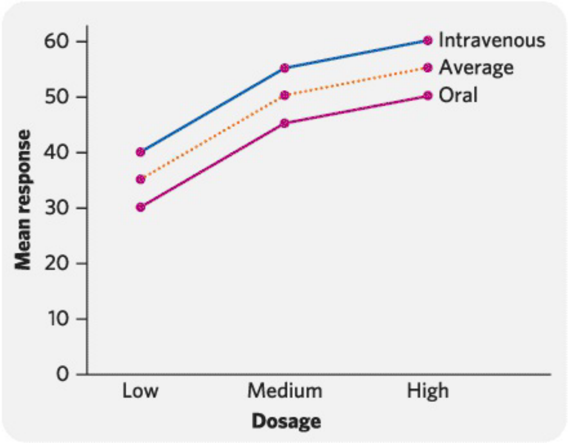


Figure 26.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 26.4 Plot of the means for the made-up data from Example 26.8. In addition to the means for each of the six conditions, the plot displays the average for each of the low, medium, and high dosage groups as a dotted line. The parallel lines show that there is no interaction between the two factors.

To see the main effect of drug dosage, look at the average response for oral and intravenous drug delivery at each dosage. This average is the dotted line in the plot. It changes as we move from low to medium to high dosages. *A variable has a main effect when the average response differs for different values of the variable.* "Average" here means averaged over all the values of the other variable. A main effect of drug dosage is present in Figure 26.4 because the dotted "average" line is not flat. A main effect for drug delivery method is also present but can't be seen directly in this plot.

EXAMPLE 26.9 Second scenario: interactions and main effects

We continue exploring the hypothetical case of drug efficacy from Example 26.7. This time, however, let's imagine that the population means are as follows:

		Variable *C*		
		Dosage		
		Low	**Medium**	**High**
Variable *R*	**Oral**	30	45	50
Delivery	**Intravenous**	40	55	60

Figure 26.5 plots these means, with the means for oral delivery connected by solid black lines and those for intravenous delivery by solid red lines.

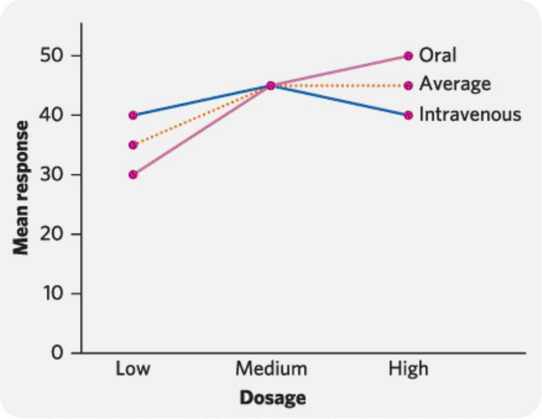


Figure 26.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 26.5 Plot of the means for the made-up data from Example 26.9, along with the averages for each of the low, medium, and high dosage groups. The lines are not parallel, so there is an interaction between delivery method and dosage.

These means reflect a situation in which response to treatment decreases when the drug is given intravenously at a high dosage. This could happen, for instance, if treatment required the drug to stay in the system as long as possible but the intravenous delivery simply “dumped” the drug in the blood all at once. The mean responses for oral delivery are the same as in Example 26.8 and Figure 26.4. But when given intravenously, the medium dosage increases the mean response by only 5 points, and the high dosage drops the mean back to 40. There *is* an interaction between delivery and dosage: The difference between oral and intravenous delivery changes with the drug dosage given, so that the solid black and red lines in Figure 26.5 are *not parallel*.

There is still a main effect of dosage, because the average response (dotted line in Figure 26.5) changes as we move from low to medium to high dosages. What about the effect of delivery method? The average over all values of dosage for the oral delivery is

30+45+503=1253=41.7

$$\frac{30+45+50}{3}=\frac{125}{3}=41.7$$

For the intravenous delivery this average is the same,

40+45+403=1253=41.7

$$\frac{40+45+40}{3}=\frac{125}{3}=41.7$$

On average over all drug dosages, changing the delivery method has no effect. There is *no main effect* for delivery method. The interaction blurred the overall, “on average” effect of delivery method.

INTERACTIONS AND MAIN EFFECTS

An **interaction** is present between factors *R* and *C* in a two-way layout if the change in mean response when we move between two values of *R* is different for different values of *C*.

(We can interchange the roles of R and C in this statement.)

A **main effect** for factor R is present if, when we average the responses for a fixed value of R over all values of C, we do not get the same result for all values of R.

CAUTION

Main effects may have little meaning when interaction is present. After all, interaction says that the effect of changing one of the variables is different for different values of the other variable. The main effect, as an "on average" effect, may not tell us much. In Example 26.9, there is no main effect for the drug delivery method. But delivery method certainly matters—Figure 26.5 shows that there is little point giving the high drug dosage to patients if they will receive it intravenously. That's the interaction of dosage with delivery method.

EXAMPLE 26.10 Which effect is most important?

There are no simple rules for interpreting results from two-way ANOVA when strong interaction is present. You must look at plots of means and think. Figure 26.6 displays two different mean plots for a made-up study of the effects of classroom conditions on the performance of normal and hyperactive schoolchildren. The two conditions are "quiet" and "noisy," where the noisy condition is actually the usual environment in elementary school classrooms.

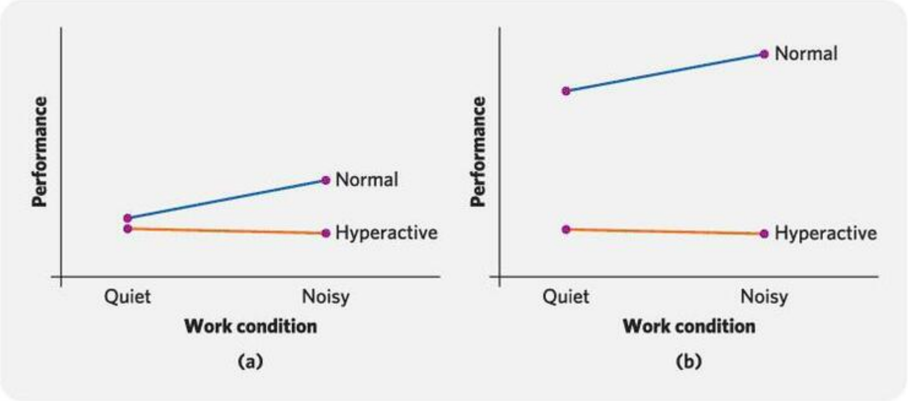


Figure 26.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.6 Plots of two sets of means for a made-up study comparing the performance of normal and hyperactive children under two conditions, for Example 26.10. The lines in

(a) are not parallel, indicating an interaction between the two factors. In addition to the interaction effect, the large gap between the two lines in (b) indicates a very strong main effect of whether a child is hyperactive or not.

There is an interaction: Typical ("normal") children perform a bit better under noisy conditions, but hyperactive children perform slightly less well under noisy conditions. The interaction is exactly the same size in the two plots of Figure 26.6. To see this, look at the slopes of the "Hyperactive" lines in the two plots: They are the same. The slopes of the two "Normal" lines are also the same. So the size of the gap between normal and hyperactive changes by the same amount when we move from quiet to noisy in both plots even though the gap is much larger in Figure 26.6(b).

In Figure 26.6(a), this interaction is the most important conclusion of the study. Both main effects are small: Normal children do a bit better than hyperactive children, for example, but not a great deal better on the average.

In Figure 26.6(b), the main effect of "hyperactive or not" is the big story. Normal children perform much better than hyperactive children in both environments. The interaction is still there, but it is not very important in the face of the large difference in average performance between hyperactive and normal children.

In this section we pretended that we knew the population means so that we could discuss patterns without needing statistical inference. In practice, we simply don't know the population means. However, plotting the sample means for all groups is an essential part of data analysis for a two-way layout. Examine the graph of sample means for interaction and main effects just as we did in this section. Of course, you will almost never find exactly parallel lines representing exactly no interaction in real data. Two-way ANOVA inference helps guide you because it assesses whether the interaction in the data is statistically significant. We are now ready to introduce two-way ANOVA inference.

APPLY YOUR KNOWLEDGE

Figure 26.7 shows two made-up plots of means for a two-way study that compares the yields of two varieties of soybeans (V1 and V2) when three different herbicides (H1, H2, and H3) are applied to the fields. Exercises 26.7 and 26.8 ask you to interpret these plots.

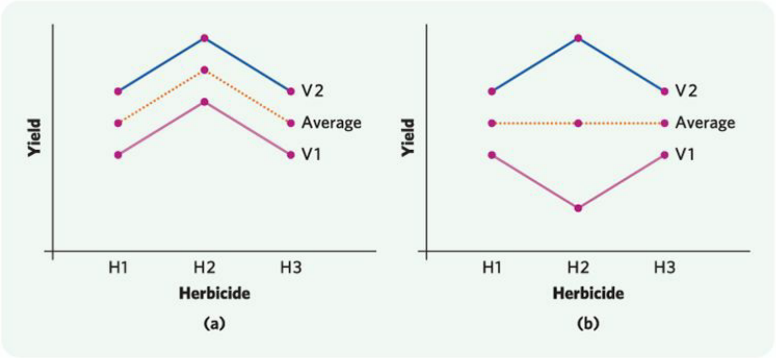


Figure 26.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.7 Plots of two sets (a and b) of possible means (solid lines) for a study in which the two factors are soybean variety and type of herbicide. The plots also show the average for each herbicide (dotted line).

26.7 Recognizing effects. Consider the mean responses plotted in Figure 26.7(a).

a. Is there an interaction between soybean variety and herbicide type? Why or why not?
b. Is there a main effect of herbicide type? Why or why not?
c. Is there a main effect of soybean variety? Why or why not?

26.8 Recognizing effects. Consider the mean responses plotted in Figure 26.7(b).

a. Is there an interaction between soybean variety and herbicide type? Why or why not?
b. Is there a main effect of herbicide type? Why or why not?
c. Is there a main effect of soybean variety? Why or why not?

26.9 Mycorrhizal symbiosis and salt stress. Mycorrhizal symbiosis can help plants alleviate salt stress. Researchers compared the growth of mycorrhizal and nonmycorrhizal lettuce plants grown from seeds under 3 saline watering conditions (0, 40, and 80 millimoles

NaCl/liter, mmol/l). The mean shoot dry weights (in grams, g) after 7 weeks are displayed below, based on 5 lettuce plants grown in each condition:[6]

	Salinity (in mmol/l)		
	0	**40**	**80**
Nonmycorrhizal	1.34	0.57	0.42
Mycorrhizal	2.54	1.77	1.51

Plot the means, and discuss the interaction and the two main effects.

INFERENCE FOR TWO-WAY ANOVA

Inference for two-way ANOVA is in many ways similar to inference for one-way ANOVA. Here is a brief outline:

1. Find and plot the group sample means. Study the plot to understand the interaction and main effects. Do data analysis to check the conditions for ANOVA.
2. Use software for basic ANOVA inference. There are now three *F* tests with three *P*-values, which answer the questions
   Is the interaction statistically significant?
   Is the main effect for variable *R* statistically significant?
   Is the main effect for variable *C* statistically significant?
3. You may wish to carry out a follow-up analysis. For example, Tukey's method makes pairwise comparisons among the means of all treatment groups.

We will illustrate two-way ANOVA inference with several examples. In the first example, the interaction is both small and insignificant, so that the message is in the main effects.

EXAMPLE 26.11 Dietary manipulations in fruit flies

STATE: Reproduction has a high physiological cost. A diet rich in proteins can trigger increased reproductive output in fruit flies, which we would expect to lead to the depletion of reserves such as body fat. An experiment assessed the percent of body fat in female fruit flies fed one of four diets, three of which were enriched with yeast (a high-protein food). The experiment used both wild-type fruit flies and mutants with a longer reproductive cycle. There are 8 groups in a two-way layout:

	Amount of yeast in diet (mg)			
	0	**1**	**3**	**7**
Wild-type	Group 1	Group 2	Group 3	Group 4
Mutant	Group 5	Group 6	Group 7	Group 8

Martin Shields/Science Source

The two factors are the genotype (wild or mutant) and the amount of yeast in the diet (in milligrams per day). Table 26.1 displays the data.[7] The response variable is the percent of body fat (lipid) after two weeks on the diet.

TABLE 26.1 Percent lipid in female fruit flies by genotype and diet

Yeast amount (mg)	Fruit fly genotype				
			Mutant		
0	25.87	26.33	22.75	22.90	23.09
1	23.87	23.35	20.71	27.15	21.45
3	18.59	17.16	11.09	13.69	12.46
7	10.57	16.59	12.50	14.90	10.42
			Wild-type		
0	23.99	27.54	26.23	24.18	26.04
1	23.01	20.35	22.66	24.70	21.02
3	15.17	15.47	15.98	14.83	14.14
7	11.25	12.87	11.31	12.29	10.11

PLAN: Plot the sample means and discuss interaction and main effects. Check the conditions for ANOVA inference. Use two-way ANOVA F tests to determine the significance of interaction and main effects.

SOLVE: The experiment has a randomized block design, with fruit flies randomly assigned to diets separately for each genotype. We consider the blocks (genotypes) as one of the factors in analyzing the data because we are interested in comparing them.

Figure 26.8 displays side-by-side dotplots of the percent of body fat for the 5 observations in each group. The dotplots show no departures from Normality. The group standard deviations do not satisfy our rule of thumb that the largest (3.17) be no more than twice the smallest (0.69). *As the number of treatment groups increases, even samples from populations with exactly the same standard deviation are more likely to produce sample standard deviations that violate our "twice as large" rule of thumb.* (Think of comparing the shortest and tallest person among more and more people.) So our rule of thumb is often conservative for two-way ANOVA. Looking back at the dotplots, we can see that the data all cover a similar range and do not give the clear impression that they might have come from populations with different standard deviations. In fact, the smallest sample standard deviation arose from a sample with closely clustered observations. This chance event can easily happen with very small sample sizes, as in this experiment. We will proceed with ANOVA inference assuming that the population variances are similar enough, but we should keep in mind that our calculations could be somewhat inaccurate if this assumption is unfounded.

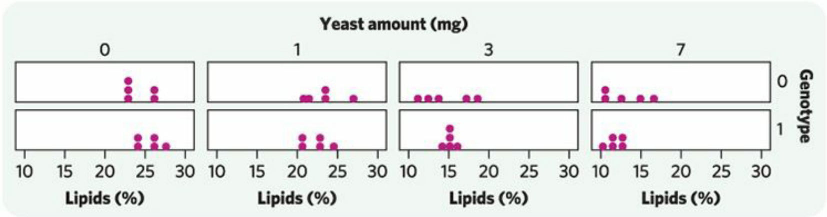


Figure 26.8

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.8 Side-by-side dotplots comparing the percent of body fat of wild-type and mutant female fruit flies for four different amounts of yeast in the diet, for Example 26.11. Genotype 0 = mutant; Genotype 1 = wild-type.

Figure 26.9 shows the plots of means produced by Minitab. The two plots display the same 8 sample means, with each variable marked on the horizontal axis. The plots are easy to interpret: The interaction and the main effect of genotype are both small, and the main effect of yeast amount is quite large. Figure 26.10 shows the two-way ANOVA output from Minitab. The three *F* tests in the Minitab output substantiate what the plots of means show: Interaction (P=0.395)$(P = 0.395)$ and genotype (P=0.860)$(P = 0.860)$ are not significant, but yeast amount (P<0.001)

$(P < 0.001)$ is highly significant.

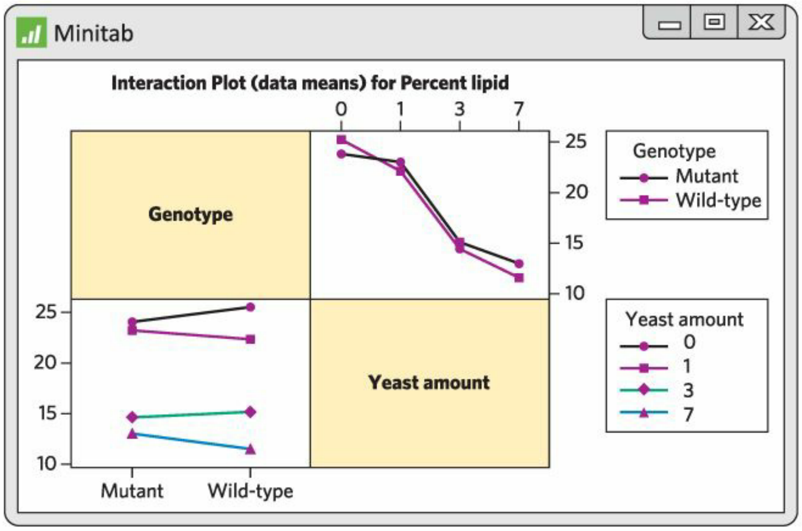
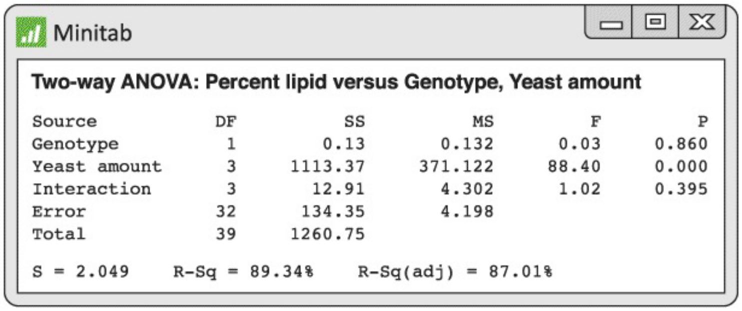


Figure 26.9

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.9 Plots of the group means from the fruit fly study, from Minitab. The two plots use the same eight means. They differ only in the choice of which variable to mark on the horizontal axis.



Two-way ANOVA: Percent lipid versus Genotype, Yeast amount

```
Source          DF        SS        MS       F      P
Genotype         1      0.13     0.132    0.03  0.860
Yeast amount     3   1113.37   371.122   88.40  0.000
Interaction      3     12.91     4.302    1.02  0.395
Error           32    134.35     4.198
Total           39   1260.75

S = 2.049   R-Sq = 89.34%   R-Sq(adj) = 87.01%
```

Figure 26.10

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.10 Partial two-way ANOVA output from Minitab for the fruit fly study data in Table 26.1.

CONCLUDE: Higher amounts of protein in the diet, in the form of yeast supplements, lead to a depletion of fat reserves in fruit flies (other results in the study link this effect to an increased reproductive output). This is the only significant effect that appears in these data. In particular, the genotype of the fruit flies has very little effect on body fat at any amount of dietary yeast.

The next example illustrates the situation in which there is significant interaction, but main effects are larger and more important. Think of Figure 26.6(b).

EXAMPLE 26.12 Mycorrhizal colonies and plant nutrition

STATE: Mycorrhizal fungi are present in the roots of many plants. This is a symbiotic relationship, in which the plant supplies nutrition to the fungus and the fungus helps the plant absorb nutrients from the soil. An experiment compared the effects of adding nitrogen fertilizer to two genotypes of tomato plants, a wild-type variety susceptible to mycorrhizal colonies and a mutant variety that is not. Nitrogen was added at rates of 0, 28, or 160 kilograms per hectare (kg/ha). Here is the two-way layout for the 6 treatment combinations:

		Tomato genotype	
		Mutant	**Wild**
	0 kg/ha	Group 1	Group 4
Nitrogen	**28 kg/ha**	Group 2	Group 5
	160 kg/ha	Group 3	Group 6

Six plants of each type were assigned at random to each level of fertilizer. The response variables describe the level of nutrients in a plant after 19 weeks, when the tomatoes are fully ripe. We will look at one response, the percent of phosphorus in the plant. Table 26.2 contains the data.[8]

TABLE 26.2 Percent of phosphorus in tomato plants

Group	Genotype	Nitrogen amount (kg/ha)	Percent phosphorus	Group	Genotype	Nitrogen amount (kg/ha)	Perce phosph
1	mutant	0	0.29	4	wild-type	0	0.64
1	mutant	0	0.25	4	wild-type	0	0.54
1	mutant	0	0.27	4	wild-type	0	0.53
1	mutant	0	0.24	4	wild-type	0	0.52

1	mutant	0	0.24	4	wild-type	0	0.41
1	mutant	0	0.20	4	wild-type	0	0.43
2	mutant	28	0.21	5	wild-type	28	0.41
2	mutant	28	0.24	5	wild-type	28	0.37
2	mutant	28	0.21	5	wild-type	28	0.50
2	mutant	28	0.22	5	wild-type	28	0.43
2	mutant	28	0.19	5	wild-type	28	0.39
2	mutant	28	0.17	5	wild-type	28	0.44
3	mutant	160	0.18	6	wild-type	160	0.34
3	mutant	160	0.20	6	wild-type	160	0.31
3	mutant	160	0.19	6	wild-type	160	0.36
3	mutant	160	0.19	6	wild-type	160	0.37
3	mutant	160	0.16	6	wild-type	160	0.26
3	mutant	160	0.17	6	wild-type	160	0.27

PLAN: Plot the sample means and discuss interaction and main effects. Check the conditions for ANOVA inference. Use two-way ANOVA *F* tests to determine the significance of interaction and main effects.

SOLVE: The design is a randomized block design, with tomatoes randomly assigned to nitrogen level separately for each genotype. We consider the blocks (genotypes) as one of the factors in analyzing the data because we are interested in comparing the two genotypes.

Figure 26.11 displays Minitab’s plots of the 6 sample means. The lines are not parallel, so interaction is present. The interaction is rather small compared with the main effects. The main effect of genotype is expected: Wild-type tomato plants with mycorrhizal colonies have higher phosphorus levels than the mutants at all levels of fertilization, because they benefit from symbiosis with the fungus. The main effect of fertilizer is a bit surprising: Phosphorus level goes down as the level of nitrogen fertilizer increases.

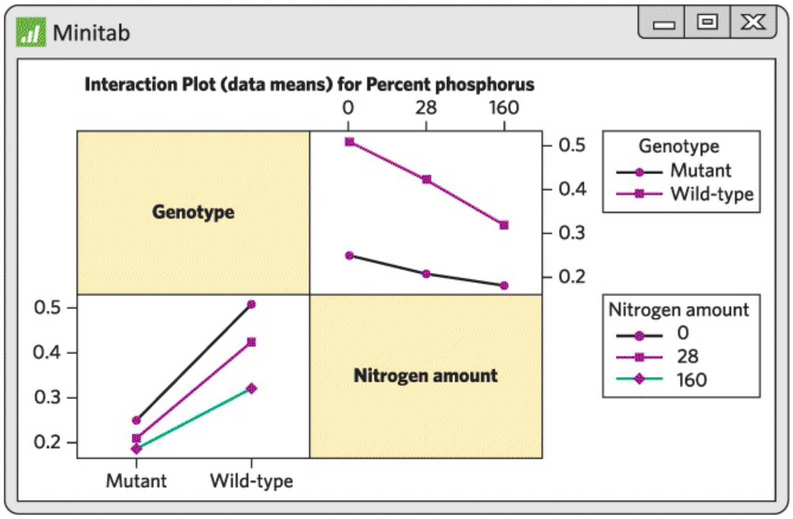


Figure 26.11

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.11 Plots from Minitab of the group means for the study of phosphorus levels in tomatoes, for Example 26.12.

Examination of the data (we don't show the details) finds no outliers or strong skewness. But the largest sample standard deviation (0.083 in Group 4) is much larger than twice the smallest (0.014 in Group 3). As in Example 26.11, it is not entirely unexpected to find relatively large differences in sample standard deviations when the sample sizes are small and the number of treatment groups is large. Nonetheless, ANOVA inference may not give correct *P*-values for these data. The *P*-values for the three two-way ANOVA *F* tests are P=0.008$P = 0.008$ for interaction and P<0.001$P < 0.001$ for both main effects. These agree with the mean plots and are so small that even if not accurate, they strongly suggest significance.

CONCLUDE: Wild-type plants, with their mycorrhizal colonies, have higher phosphorus levels than mutants that lack such colonies. Nitrogen fertilizer actually reduces phosphorus levels in both types of plants. The reduction is stronger for wild-type plants, but this interaction is not very large in practical terms.

Finally, here is an example in which strong interaction makes one of the main effects meaningless. Two-way ANOVA with strong interaction is often difficult to interpret simply, as the following example illustrates.

EXAMPLE 26.13 Better corn for heavier chicks?

STATE: Corn varieties with altered amino acid content might be advantageous for feeding animals. Here is an excerpt from a study that compared normal corn (“norm” in the data file *eg26-13.dat*) with two altered varieties called opaque-2 (“opaq”) and floury-2 (“flou”).

> Nine treatments were arranged in a 3×33 × 3 factorial experiment to compare opaque-2, floury-2, and normal corn at dietary protein levels of 20, 16, and 12%. Corn-soybean meal diets containing either opaque-2, floury-2, or normal corn were formulated so that, for a given protein level, an equivalent amount of corn protein was supplied by each type of corn. Male broiler-type chicks were randomly allotted to treatments at 1 day of age. Feed and water were provided *ad libitum*. Chicks were weighed at weekly intervals until termination of the experiment at 21 days.[9]

There are 10 chicks in each group. The response variable is the weight in grams after 21 days. Which combinations of corn type and protein content lead to the most growth?

PLAN: Plot the sample means and discuss interaction and main effects. Check the conditions for ANOVA inference. Use two-way ANOVA *F* tests to determine the significance of interaction and main effects. If necessary, use Tukey pairwise comparisons to identify significant differences among treatments.

SOLVE: The experiment has a balanced, completely randomized design with 9 treatments. Figure 26.12 shows Minitab’s plots of the sample means. The mean weight of the chicks increases with the percent of protein in the diet, as expected. We are primarily interested in comparing the three types of corn. There are important interaction effects. Normal corn does poorest of the three types at 12% protein and best at 20%. Floury is best at both 12% and 16%. Opaque is always inferior to floury and beats normal corn only at 12% protein.

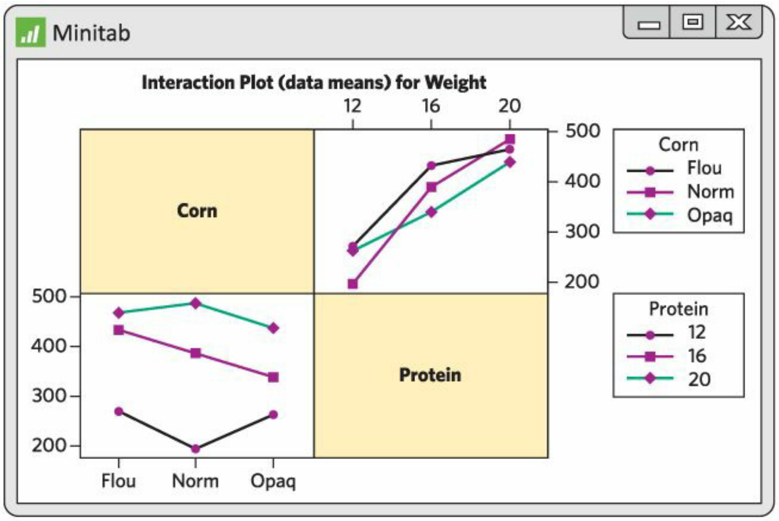
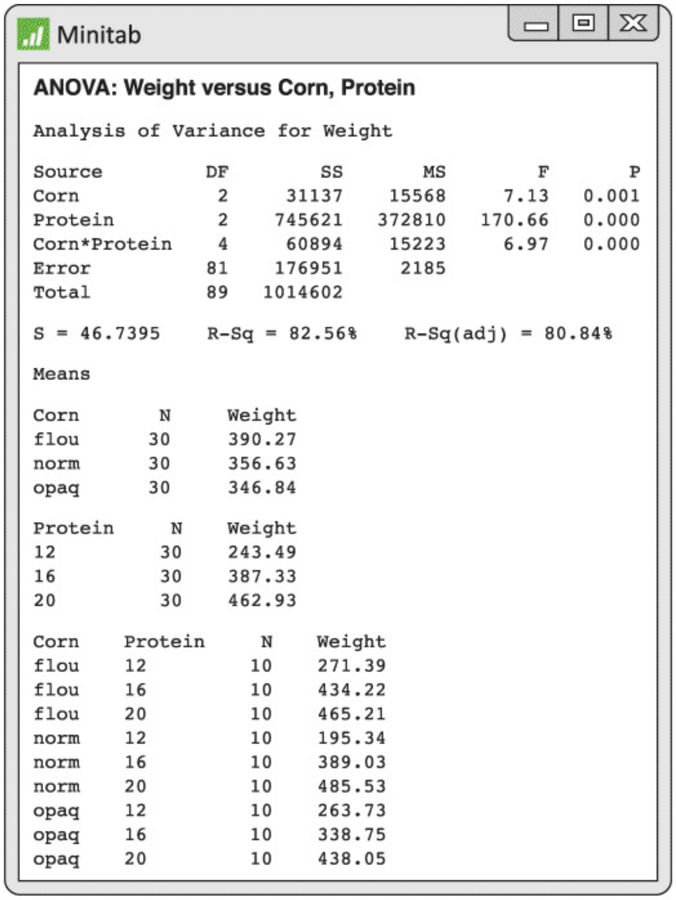


Figure 26.12

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.12 Plots from Minitab of the mean weights of 21-day-old chicks fed 9 different diets, for Example 26.13.

Although we don't show the details, ANOVA is justified. There are no outliers or strong skewness. The largest sample standard deviation (61.16 for Group 4) is a little bit more than twice the smallest (25.99 for Group 6), but this is common when we have 9 groups even when all populations really have the same σ.σ. Figure 26.13 contains Minitab's ANOVA output. We included the means for the 9 groups, for the three types of corn, and for the three levels of protein. The main effects can be seen in the different mean weights for the corn types and for the protein levels. The three *F* tests are all highly significant. Although there is a substantial main effect for corn (the means range from 346.84 g for opaque to 390.27 g for floury), this has little meaning in light of the interaction that we just described.



```
ANOVA: Weight versus Corn, Protein

Analysis of Variance for Weight

Source        DF       SS      MS       F      P
Corn           2    31137   15568    7.13  0.001
Protein        2   745621  372810  170.66  0.000
Corn*Protein   4    60894   15223    6.97  0.000
Error         81   176951    2185
Total         89  1014602

S = 46.7395   R-Sq = 82.56%   R-Sq(adj) = 80.84%

Means

Corn     N   Weight
flou    30   390.27
norm    30   356.63
opaq    30   346.84

Protein    N   Weight
12        30   243.49
16        30   387.33
20        30   462.93

Corn   Protein     N   Weight
flou   12         10   271.39
flou   16         10   434.22
flou   20         10   465.21
norm   12         10   195.34
norm   16         10   389.03
norm   20         10   485.53
opaq   12         10   263.73
opaq   16         10   338.75
opaq   20         10   438.05
```

Figure 26.13
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 26.13 Two-way ANOVA output from Minitab for the study of the effect of diet on chick growth, for Example 26.13.

When strong interaction makes one or both main effects hard to interpret, it is often useful to find the Tukey pairwise multiple comparisons for all the treatment groups. One way to do this is to do a one-way ANOVA with just “group” as the explanatory variable. There are 36 pairwise comparisons among 9 groups. Minitab’s Tukey output is both long and hard to understand in such cases. Here is a condensed version using an idea that some software (though not Minitab) implements. Arrange the 9 groups in the order of their sample means, from smallest to largest. We have identified the groups both by their group number in the data file and by the treatment. Connect all pairs that do *not* differ significantly at the overall 5% level with an underline:

```
Treatment:  N12  O12  F12  O16  N16  F16  O20  F20  N20
 Group:      4    7    1    8    5    2    9    3    6
                  ------    ------
                                 -----------
                                      ----------------
```

Group 4 has a significantly smaller mean weight than any other group. Groups 7 and 1 do not differ significantly, but they are higher than Group 4 and lower than all other groups. Group 8 is not significantly different from 5 but is higher than 4, 7, and 1 and lower than 2, 9, 3, and 6. And so on. The most interesting finding is that at the high end, Groups 2, 9, 3, and 6 do not have significantly different mean weights. Three of these are the three 20% protein groups, but floury corn with 16% protein belongs with these three.

CONCLUDE: More protein clearly helps chicks grow faster. The three types of corn do not differ significantly when the diet has 20% protein. Floury corn is superior to both opaque and normal corn at middle (16%) and low (12%) protein levels, though not all differences at these levels are statistically significant.

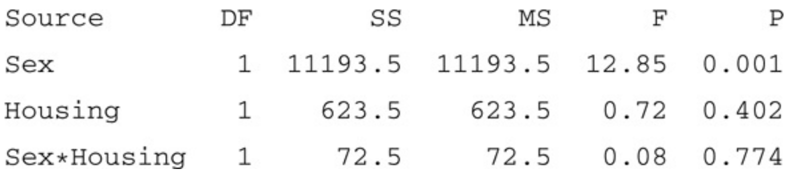
APPLY YOUR KNOWLEDGE

26.10 Hooded rats: social play times. How does social isolation during a critical development period affect the behavior of hooded rats? Psychology students assigned 24 young female rats at random to either isolated or group housing, then similarly assigned 24 young male rats. This is a randomized block design with the gender of the 48 rats as the blocking variable and housing type as the treatment. Later, the students observed the rats at play in a group setting and recorded data on three types of behavior (object play, locomotor play, and social play).[10] The file *ex26-10.dat* records the time (in seconds) that each rat devoted to social play during the observation period.

a. Make a plot of the 4 group means. Is there a large interaction between gender and housing type? Which main effect appears to be more important?

b. Verify that the conditions for ANOVA inference are satisfied.

c. Here is the ANOVA table from Minitab:



```
Source         DF        SS        MS      F      P
Sex             1   11193.5   11193.5  12.85  0.001
Housing         1     623.5     623.5   0.72  0.402
Sex*Housing     1      72.5      72.5   0.08  0.774
```

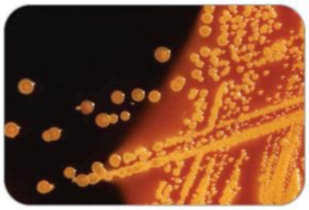
NHPA/SuperStock

What are the *F* statistics and *P*-values for interaction and the two main effects? Do these values support your tentative interpretation of the graph?

26.11 Evolution in bacteria. Exercise 24.31 (page 630) described an experiment on *E. coli* bacteria to test the theory of evolution. Lines of *E. coli* bacteria grown and kept at a neutral pH of 7.2 were “evolved,” that is, grown for 2000 generations (about 300 days) at a stressful acidic pH of 5.5 and then tested against their ancestor at various pH values to determine their relative fitness. A control was also performed in which bacteria were “evolved” at the neutral pH and tested against their ancestors at various pH values. The objective was to determine whether any change in relative fitness could be attributed to directional evolution (as a result of an environmental change) or simply a chance event. Here are the relative fitness scores obtained for test environments of pH 5.5 (acid), 7.2 (neutral), or 8.0 (basic). There were six replicates using different original bacterial lines for each test group:[11]

Bacteria “evolved” at pH 5.5						
Test in acid pH	1.24	1.22	1.23	1.24	1.18	1.09
Test in neutral pH	0.99	0.99	0.98	0.94	0.95	0.95
Test in basic pH	0.56	0.83	0.82	0.72	0.86	0.84
Bacteria “evolved” at pH 7.2						

Test in acid pH	1.02	1.04	0.99	1.10	1.12	1.08
Test in neutral pH	1.15	1.06	1.04	0.93	1.02	1.03
Test in basic pH	0.80	1.06	1.07	1.11	0.96	1.04

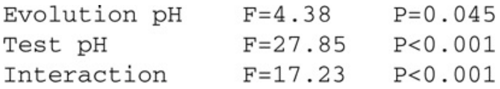


CDC

A score of 1 indicates that both bacteria types (evolved and ancestral) are equally fit. A score larger than 1 indicates that the evolved line is more fit than the ancestral line.

a. Plot the means on a single graph and describe the main effects and interaction.

b. Software gives the following information:



```
Evolution pH    F=4.38    P=0.045
Test pH         F=27.85   P<0.001
Interaction     F=17.23   P<0.001
```

Do these values support your interpretation of the graph?

c. Check that the conditions for ANOVA are met. If you find deviations from the recommendations, describe them and explain how they might affect your interpretation in b.

SOME DETAILS OF TWO-WAY ANOVA[ii]

All ANOVA *F* statistics work on the same principle: Compare the variation due to the effect being tested with a benchmark level of variation that would be present even if that effect were absent. The three *F* tests for two-way ANOVA use the same benchmark as the one-way ANOVA *F* test, namely, the variation among individual responses within each treatment group.

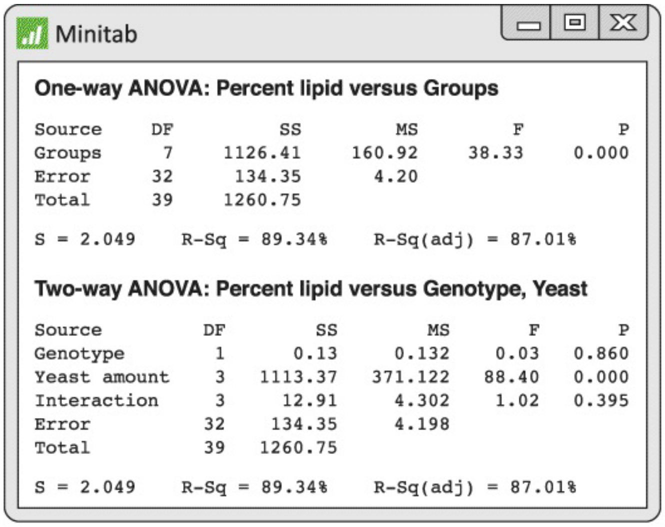
In two-way ANOVA, we have two factors (also called explanatory or independent variables) that form treatments in a two-way layout. Factor *R* has *r* values and Factor *C* has *c* values, so that there are *rc* treatments. If the design is balanced, then the same number *n* of subjects are assigned to each treatment. The two-way layout that results is as follows:

		Column factor *C*			
		1	**2**	…	***c***
	1	*n* subjects	*n* subjects	…	*n* subjects
Row	**2**	*n* subjects	*n* subjects	…	*n* subjects
factor *R*	⋮	⋮	⋮		⋮
	r	*n* subjects	*n* subjects	…	*n* subjects

The number of treatments is k=rcThe total number of observations is N=rcn=kn

The number of treatments is $k = rc$
The total number of observations is $N = rcn = kn$

Figure 26.14 presents both one-way and two-way ANOVA output for the same data, from the fruit fly study in Example 26.11. The one-way analysis just compares the means of rcrc treatments, ignoring the two-way layout. In discussing one-way ANOVA, we called the number of treatments *k*. Now k=rc.$k = rc.$ The two-way analysis takes into account that each treatment is formed by combining a value of *R* with a value of C.c



```
One-way ANOVA: Percent lipid versus Groups

Source    DF       SS      MS      F      P
Groups     7  1126.41  160.92  38.33  0.000
Error     32   134.35    4.20
Total     39  1260.75

S = 2.049   R-Sq = 89.34%   R-Sq(adj) = 87.01%

Two-way ANOVA: Percent lipid versus Genotype, Yeast

Source         DF       SS       MS      F      P
Genotype        1     0.13    0.132   0.03  0.860
Yeast amount    3  1113.37  371.122  88.40  0.000
Interaction     3    12.91    4.302   1.02  0.395
Error          32   134.35    4.198
Total          39  1260.75

S = 2.049   R-Sq = 89.34%   R-Sq(adj) = 87.01%
```

Figure 26.14
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 26.14 Compare the sums of squares in these one-way and two-way ANOVA outputs for the fruit fly data from Table 26.1.

In both settings, analysis of variance breaks down the overall variation in the observations into several pieces. The overall variation is expressed numerically by the **total sum of squares**

SST=∑(individual observation - mean of all observations)2

$$\text{SST} = \sum (\text{individual observation} - \text{mean of all observations})^2$$

where the sum runs over all NN individual observations. If we divide SST by N-1,$N - 1,$ we get the variance of the observations. So SST is closely related to a familiar measure of variability. Because SST uses just the NN individual observations, it is the same for both one-way and two-way analyses. You can see in Figure 26.14 that SST = 1260.75 for these data.

One-way ANOVA. We saw in Chapter 24 that the one-way ANOVA *F* test compares the variation among the *k* treatment means with the variation among responses to each treatment. If the means vary more than we would expect based on the variation among subjects who receive the same treatment, that's evidence of a difference among the mean responses in the *k* populations.

Let's give a bit more detail. One-way ANOVA breaks down the total variation into the sum of two parts:

total variation among responses=variation among treatment means+variation among responses to each treatmenttotal sum of squares=sum of squares for groups+sum of squares for errorSST=SSG+SSE

$$\begin{aligned}\text{total variation among responses} &= \text{variation among treatment means}\\ &+ \text{variation among responses to each treatment}\\ \text{total sum of squares} &= \text{sum of squares for groups}\\ &+ \text{sum of squares for error}\\ \text{SST} &= \text{SSG} + \text{SSE}\end{aligned}$$

Formulas for the sum of squares for groups (SSG) and the sum of squares for error (SSE) appear in Chapter 24 as the numerators of the mean square for groups (MSG) and the mean square for error (MSE), respectively, but we won't concern ourselves with the algebra. Remember that "error" is the traditional term in ANOVA for variation among observations. It doesn't imply that some mistake has been made. In the one-way output in Figure 26.14, you see that the breakdown for these data is, aside from rounding error,

SST=SSG+SSE1260.75=1126.41+134.35

$$\begin{aligned}\text{SST} &= \text{SSG} + \text{SSE}\\ 1260.75 &= 1126.41 + 134.35\end{aligned}$$

The one-way ANOVA *F* test is performed in two stages:

1. Divide each sum of squares by its *degrees of freedom* to get the *mean squares* MSG for groups and MSE for error:

   MSG=SSGk-1   MSE=SSEN-k

   $$\text{MSG} = \frac{\text{SSG}}{k-1} \qquad \text{MSE} = \frac{\text{SSE}}{N-k}$$

2. The one-way ANOVA *F* statistic compares MSG with MSE:

   F=MSGMSE

   $$F = \frac{\text{MSG}}{\text{MSE}}$$

   Find the *P*-value from the *F* distribution with k-1$k-1$ and N-k$N-k$ degrees of freedom.

The **analysis of variance table** in the output reports sums of squares, their degrees of freedom, mean squares, and the *F* statistic with its *P*-value.

Two-way ANOVA. Now look at the two-way analysis of variance table in Figure 26.14:

- The total sum of squares and the error sum of squares are the same as in the one-way analysis.
- The sum of squares for groups in one-way is the sum of the three sums of squares for main effects and interaction in two-way.

This is the heart of two-way analysis of variance: Break down the variation among the *rc* groups into three parts—variation due to the main effect of Factor *R*, variation due to the main effect of Factor *C*, and variation due to interaction between the two factors. Each type of variation is measured by a sum of squares. The formulas for the two main effects sums of squares are similar to that for the one-way sum of squares for groups, but we will again ignore the algebraic details. The interaction sum of squares is best thought of as what's left over, the variation among treatments that isn't explained by the two main effects. In symbols,

total sum of squares=sum of squares for main effect of Factor R+sum of squares for main effect of Factor C+sum of squares for interaction between R and C+sum of squares for errorSST=SSR+SSC+SSRC+SSE

$$\begin{aligned} \text{total sum of squares} &= \text{sum of squares for main effect of Factor } R \\ &+ \text{sum of squares for main effect of Factor } C \\ &+ \text{sum of squares for interaction between } R \text{ and } C \\ &+ \text{sum of squares for error} \\ \text{SST} &= \text{SSR} + \text{SSC} + \text{SSRC} + \text{SSE} \end{aligned}$$

Each of these sums of squares has its own degrees of freedom, and these break down in the same way:

total df=df for main effect of Factor R+df for main effect of Factor C+df for interaction between R and C+df for errorrcn-1=(r-1)+(c-1)+(r-1)(c-1)+rc(n-1)=N-1

$$\begin{aligned} \text{total df} &= \text{df for main effect of Factor } R \\ &+ \text{df for main effect of Factor } C \\ &+ \text{df for interaction between } R \text{ and } C \\ &+ \text{df for error} \\ rcn-1 &= (r-1)+(c-1)+(r-1)(c-1)+rc(n-1) \\ &= N-1 \end{aligned}$$

You can check that the total degrees of freedom are N-1$N-1$ and the degrees of freedom for error are N-k,$N-k,$ the same as for one-way ANOVA. The one-way degrees of freedom for groups are equal to the sum of the degrees of freedom for the three two-way effects.

EXAMPLE 26.14 Comparing one-way and two-way ANOVA

The data behind Figure 26.14 appear in Table 26.1. The two factors are R=genotype$R = \text{genotype}$ and C=yeast amount.$C = \text{yeast amount}.$ Factor *R* has r=2$r = 2$ values: mutant and wild-type. Factor *C* has c=4$c = 4$ values: 0 mg, 1 mg, 3 mg, and 7 mg. There are k=8$k = 8$ treatments and n=5$n = 5$ observations for each treatment, resulting in N=40$N = 40$ observations.

The total degrees of freedom are N-1=39.$N - 1 = 39.$ The degrees of freedom for error is N-k=40-8=32.$N - k = 40 - 8 = 32.$ In the one-way analysis, the degrees of freedom for “Groups” are k-1=7.$k - 1 = 7.$ The two-way analysis breaks this into degrees of freedom r-1=1$r - 1 = 1$ for Factor R,$R,$ c-1=3$c - 1 = 3$ for Factor C,$C,$ and (r-1)(c-1)=1×3=3 $(r - 1)(c - 1) = 1 \times 3 = 3$ for interaction.

Here are the two breakdowns of the total variation and the degrees of freedom that appear in Figure 26.14:

One-way			**Two-way**		
Sums of squares		**df**	**Sums of squares**		**df**
SSG	1126.41	7	SSR	0.13	1
			SSC	1113.37	3
			SSRC	12.91	3
SSE	134.35	32	SSE	134.35	32
SST	1260.75	39	SST	1260.75	39

The neat breakdown of SSG into three effects depends on the balance of the two-way layout. It doesn’t hold when the counts of observations are not the same for all treatments. That’s why two-way ANOVA is more complicated and harder to interpret for unbalanced data than it is for balanced data.

Two-way ANOVA *F* tests. Finally, form three *F* statistics exactly as in the one-way setting.

1. Divide each sum of squares by its *degrees of freedom* to get the *mean squares* for the three effects and for error:

MSR=SSRr-1 MSC=SSCc-1 MSRC=SSRC(r-1)(c-1) MSE=SSEN-k

$$\text{MSR} = \frac{\text{SSR}}{r-1} \qquad \text{MSC} = \frac{\text{SSC}}{c-1} \qquad \text{MSRC} = \frac{\text{SSRC}}{(r-1)(c-1)}$$

$$\text{MSE} = \frac{\text{SSE}}{N-k}$$

2. The three *F* statistics compare the mean squares for the three effects with MSE.

TWO-WAY ANOVA *F* TESTS

The *F* statistics for the three types of treatment effects in two-way ANOVA are

For the main effect of Factor R,R, F=MSRMSE$F = \frac{\text{MSR}}{\text{MSE}}$ with dfs r-1$r-1$ and N-k$N-k$

For the main effect of Factor C,C, F=MSCMSE$F = \frac{\text{MSC}}{\text{MSE}}$ with dfs c-1$c-1$ and N-k$N-k$

For the interaction of RR and C,C, F=MSRCMSE$F = \frac{\text{MSRC}}{\text{MSE}}$ with dfs (r-1)(c-1)$(r-1)(c-1)$ and N-k$N-k$

In all cases, large values of *F* are evidence against the null hypothesis that the effect is not present in the populations.

APPLY YOUR KNOWLEDGE

26.12 Hooded rats: social play times, continued. Exercise 26.10 gave the two-way ANOVA table for a study of the effect of social isolation on hooded rats. The response variable is the time (in seconds) that a rat devoted to social play during an observation period. Start your work in this exercise with the two-way ANOVA table.

a. Explain how the sums of squares from the two-way ANOVA table can be combined to obtain the one-way ANOVA sum of squares for the 4 groups (SSG). What is the value of SSG?

b. Give the degrees of freedom, mean square (MSG), and *F* statistic for testing for the effect of groups in the one-way ANOVA setting.

c. Is there a significant effect of group on the amount of time spent in play? Give and interpret the *P*-value in the context of this experiment.

d. Use software to carry out one-way ANOVA of time on group. Verify that your results in parts b, c, and d agree with the software output.

26.13 Evolution in bacteria, continued. Exercise 26.11 gives the *F* statistics and *P*-values for the main effects and interaction of a two-way ANOVA of bacterial relative fitness as a function of evolution pH and test pH.

a. What are the degrees of freedom for each of the three *F* values?

b. The mean squares (MS) for evolution pH, test pH, and interaction are 0.027225, 0.172978, and 0.107033, respectively. Use this information to fill in the complete ANOVA table for the two-way ANOVA test (you can refer to Figure 26.10 for a model).

CHAPTER 26 SUMMARY

- **Two-way analysis of variance (ANOVA)** compares the means of several populations formed by combinations of two factors R and C in a **two-way layout.**
- The **conditions for ANOVA** state that we have an **independent SRS** from each population (or a completely randomized experimental design), that each population has a **Normal distribution,** and that all populations have the **same standard deviation.** In this chapter we consider only examples that satisfy the additional conditions that the design producing the data be **crossed** (all combinations of the factors are present) and **balanced** (all factor combinations are represented by the same number of individuals).
- A factor has a **main effect** if the mean responses for each value of that factor, averaged over all values of the other factor, are not the same. The two factors **interact** if the effect of moving between two values of one factor is different for different values of the other factor. **Plot the treatment mean responses** to examine main effects and interaction.
- There are three **ANOVA *F* tests:** for the null hypotheses of no main effect for Factor R, no main effect for Factor C, and no interaction between the two factors.
- **Follow-up analysis** is often helpful in both one-way and two-way ANOVA settings. **Tukey pairwise multiple comparisons** give confidence intervals for all differences between pairs of treatment means with an **overall level of confidence.** That is, we can be (say) 95% confident that *all* the intervals simultaneously capture the true population differences between means. When the data lend themselves to more specific hypotheses, contrasts of combinations of means can be created and used for inference.

STATISTICS IN SUMMARY

Here are the most important skills you should have acquired from reading this chapter.

A. Recognition

1. Recognize the two-way layout, in which we have a quantitative response to treatments formed by combinations of values of two factors.
2. Recognize when comparing mean responses to the treatments in a two-way layout is helpful in understanding data.
3. Recognize when you can use two-way ANOVA to compare means. Check the data production, the presence of outliers, and the sample standard deviations for the groups you want to compare. Look for data production designs that are crossed and balanced.

B. Interpreting Two-Way ANOVA

1. Plot the sample means for the treatments or groups. Based on your plot, describe the main effects and interaction that appear to be present.
2. Decide which effects are most important in practice. Pay particular attention to whether a large interaction makes one or both main effects less meaningful.
3. Use software to carry out two-way ANOVA inference. From the *P*-values of the three *F* tests, learn which effects are statistically significant.

C. Follow-up Analysis

1. Decide when it is helpful to know which differences among treatment means are significant.
2. Use software to carry out Tukey pairwise multiple comparisons among all the means you want to compare.
3. Understand the meaning of the overall confidence level and the

overall level of significance provided by Tukey’s method for a set of confidence intervals or a set of significance tests.

THIS CHAPTER IN CONTEXT

In Chapter 24 we described the one-way analysis of variance *F* test, a method used as a first step in comparing the means of several populations. One-way ANOVA allows us to avoid the problems associated with multiple comparisons by first asking if there is evidence that these population means are not all equal. When the ANOVA *F* test is statistically significant, we have evidence that at least one population mean in the set is significantly different from at least one other population mean. But ANOVA does not say which ones.

In this chapter we examine techniques, such as Tukey comparisons and contrasts, that allow us to compare a set of population means two at a time while maintaining a reasonable overall probability of committing a Type I error. These follow-up tests are not fundamentally different from the two-sample *t* procedures that we saw in Chapter 18 and used to perform statistical inference for the difference μ1-μ2$\mu_1 - \mu_2$ between the means of two distinct populations.

We also expand the use of analysis of variance to more complex settings. Specifically, we describe how the two-way ANOVA procedure allows comparisons of population means obtained from designs involving two different factors, or explanatory variables. The advantage of a two-way design is that it allows us to consider possible interactions between the factors. This is a concept that we will revisit in companion Chapter 28 on multiple regression.

CHECK YOUR SKILLS

26.14 To avoid the problem of multiple comparisons in an ANOVA follow-up analysis, one should

a. use contrasts for all pairwise comparisons of means.
b. use tests that guarantee a given significance level for all pairwise comparisons of means (such as Tukey tests).
c. use a series of two-sample *t* tests for all pairwise comparisons of means.

26.15 An ANOVA comparing three population means finds a significant *P*-value. With a 5% significance level, the pairwise Tukey tests find that μ1μ_1 is significantly different from μ2μ_2 but the other tests are not significant. These results

a. say that we have significant evidence that μ1≠μ2$\mu_1 \neq \mu_2$ but not enough evidence to reject the other two null hypotheses.
b. must be wrong because if μ1≠μ2$\mu_1 \neq \mu_2$ and μ2=μ3,$\mu_2 = \mu_3$, then μ1=μ3 $\mu_1 = \mu_3$ is not possible.
c. are not very convincing because they are based on multiple comparisons.

26.16 Inference about a population contrast tests

a. whether any linear combination of population means we are interested in is equal to zero.
b. whether a linear combination of population means is equal to zero if that hypothesis made sense before seeing the sample data.
c. whether one population mean is equal or not to a linear combination of the other population means in a series of experiments.

26.17 The purpose of a two-way ANOVA is to examine

a. the combined effect of two factors on a quantitative variable.
b. the variances of two populations.
c. the means of two populations.

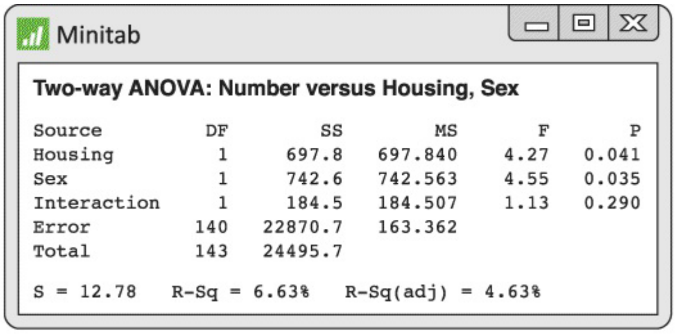
26.18 In a two-way ANOVA, an interaction effect is present when

a. the values of both factors are the same.
b. the values of both factors are different.
c. the effect of changing one of the variables is different for different values of the other variable.

26.19 In a two-way ANOVA test, when the interaction effect is significant,

a. one should interpret the main effects with caution, especially when the interaction effect is large.
b. one should simply ignore the main effects.
c. one should describe the two main effects separately.

A study examined the effects of social isolation versus social housing on the development of rats. Two-way ANOVA was used to investigate the effects of housing condition and the rats' gender on the number of play behaviors. Use the ANOVA table in Figure 26.15 to answer the questions below.



Two-way ANOVA: Number versus Housing, Sex

```
Source        DF       SS       MS      F      P
Housing        1    697.8  697.840   4.27  0.041
Sex            1    742.6  742.563   4.55  0.035
Interaction    1    184.5  184.507   1.13  0.290
Error        140  22870.7  163.362
Total        143  24495.7

S = 12.78   R-Sq = 6.63%   R-Sq(adj) = 4.63%
```

Figure 26.15
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

FIGURE 26.15 Two-way ANOVA table for Exercises 26.20 to 26.23.

26.20 How many treatment groups were there for this experiment?

a. 1
b. 2
c. 4

26.21 What is the value of the test statistic for interaction between housing condition and gender?

a. 0.041

b. 0.290
c. 1.13

26.22 Which statement below provides the best summary of the housing effect?

a. Housing is significant at the 1% level but not at the 5% level.
b. Housing is significant at the 5% level but not at the 1% level.
c. Housing is not significant at either the 1% or the 5% level.

26.23 You conclude from this experiment that the effect of sex on number of play behaviors

a. is significant at the 5% level and this effect does not depend on housing type.
b. is significant at the 5% level but this effect depends on housing type.
c. should be ignored because of the interaction effect.

CHAPTER 26 EXERCISES

26.24 Logging in the rain forest, continued. In Exercise 26.2 you compared the mean numbers of trees in forest plots in Borneo with different logging histories, using the data in Table 24.2 (page 616). This table also provides data on the variable Species, the number of tree species in a plot. The one-way ANOVA on this variable was carried out in Exercise 24.28 (page 629) and was found statistically significant (P=0.006).$(P = 0.006)$. Here are the Tukey 95% simultaneous confidence intervals given by software:

1.599to9.901forμA−μB−0.650to8.317forμA−μC−6.400to2.567forμB−μC

$$\begin{array}{rcrcl} 1.599 & \text{to} & 9.901 & \text{for} & \mu_A - \mu_B \\ -0.650 & \text{to} & 8.317 & \text{for} & \mu_A - \mu_C \\ -6.400 & \text{to} & 2.567 & \text{for} & \mu_B - \mu_C \end{array}$$

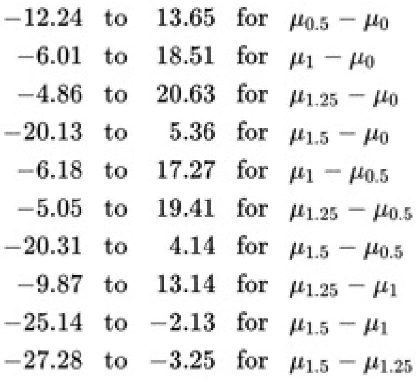
a. Write a short summary of your interpretation of the ANOVA test and follow-up analysis.

b. Use software or Table G on page 26-40 to verify the software output provided here.

26.25 Does nature heal best? The body has a natural electrical field that helps wounds heal. Might higher or lower levels speed healing? An experiment on skin healing rate in anesthetized newts compared the natural electrical field of the skin to several imposed electrical field intensities (0, 0.5, 1.25, 1.5, and the control 1). The response variable is the difference in healing rate (in micrometers per hour) of cuts made in the experimental limb and the undisturbed limb of each newt. Negative values mean that the experimental limb healed more slowly. A one-way ANOVA gives a *P*-value of 0.005. Here are the 10 pairwise Tukey 95% simultaneous confidence intervals given by software:

−12.24to13.65forμ0.5−μ0−6.01to18.51forμ1−μ0−4.86to20.63forμ1.25−μ0

−2.13forμ1.5−μ1−27.28to−3.25forμ1.5−μ1.25



-12.24	to	13.65	for	$\mu_{0.5}-\mu_0$
-6.01	to	18.51	for	$\mu_1-\mu_0$
-4.86	to	20.63	for	$\mu_{1.25}-\mu_0$
-20.13	to	5.36	for	$\mu_{1.5}-\mu_0$
-6.18	to	17.27	for	$\mu_1-\mu_{0.5}$
-5.05	to	19.41	for	$\mu_{1.25}-\mu_{0.5}$
-20.31	to	4.14	for	$\mu_{1.5}-\mu_{0.5}$
-9.87	to	13.14	for	$\mu_{1.25}-\mu_1$
-25.14	to	-2.13	for	$\mu_{1.5}-\mu_1$
-27.28	to	-3.25	for	$\mu_{1.5}-\mu_{1.25}$

a. Explain why doing a Tukey pairwise follow-up analysis is legitimate here.
b. What can you conclude from the ANOVA test and follow-up analysis?

26.26 Does nature heal best? Continued. Interpreting many pairwise Tukey comparisons can be challenging. It is always preferable to start with a specific hypothesis based on biological reasoning. Refer to Exercise 26.25, with raw data available in Table 26.3.

TABLE 26.3 **Effect of electrical field on healing rate in newts**

Level 0	Level 0.5	Level 1	Level 1.25	Level 1.5
−10	−1	−7	1	−13
−12	10	15	8	−49
−9	3	−4	−15	−16
−11	−3	−16	14	−8
−1	−31	−2	−7	−2
6	4	−13	−1	−35
−31	−12	5	11	−11
−5	−3	−4	8	−46
13	−7	−2	11	−22
−2	−10	−14	−4	2
−7	−22	5	7	10
−8	−4	11	−14	−4
	−1	10	0	−10
	−3	3	5	2
		6	−2	−5
		−1		

		13		
		−8		

a. Run a contrast test between nature's way (control group, "1") and the average of all four treatment groups. State the hypotheses, compute the statistic, and find the *P*-value. What is your conclusion?

b. Explain why this test is easier to interpret biologically than the 10 Tukey tests of Exercise 26.25.

c. Plot the means for the 5 groups on one graph. It might be tempting based on this graph to run a contrast for Groups 1 and 1.25 against all 3 other groups. Explain why this would not be appropriate.

26.27 Toxicology of lead acetate. Lead acetate is a widely used chemical often involved in dyeing textiles or hair. The National Toxicology Program (NTP) reports a study of the impact of exposure to lead acetate on the hematology of female B6C3F1 mice. The mice were randomly assigned to receive in their diet a given amount of lead acetate for 15 days. Various aspects of their hematology were examined at the end of the experiment. Here are the data for their blood hemoglobin levels after the 15-day exposure to lead acetate:[12]

Lead acetate	Hemoglobin (g/dl)							
0 (control)	16.4	16.4	16.8	16.9	17.0	16.7	17.0	16.4
500 mg/kg	15.7	16.4	16.6	16.7	17.2	17.1	15.7	16.8
1000 mg/kg	16.6	16.4	16.0	16.2	16.6	17.0	16.4	16.5
2000 mg/kg	16.0	13.5	15.1	14.4	14.9	15.0	14.0	15.3

a. Does the consumption of lead acetate influence the hemoglobin level of female B6C3F1 mice? Run the ANOVA test, following the four-step process outlined in Chapter 24.

b. Here are the 6 pairwise Tukey 95% simultaneous confidence intervals given by software for this study:

−0.9060to0.5560forμ500−μ0−0.9995to0.5138forμ1000−μ0−2.6560to
−1.1940forμ2000−μ0−0.8245to0.6888forμ1000−μ500−2.4810to
−1.0190forμ2000−μ500−2.4388to0.9255forμ2000−μ1000

$$\begin{array}{rcrcl}
-0.9060 & \text{to} & 0.5560 & \text{for} & \mu_{500} - \mu_0 \\
-0.9995 & \text{to} & 0.5138 & \text{for} & \mu_{1000} - \mu_0 \\
-2.6560 & \text{to} & -1.1940 & \text{for} & \mu_{2000} - \mu_0 \\
-0.8245 & \text{to} & 0.6888 & \text{for} & \mu_{1000} - \mu_{500} \\
-2.4810 & \text{to} & -1.0190 & \text{for} & \mu_{2000} - \mu_{500} \\
-2.4388 & \text{to} & 0.9255 & \text{for} & \mu_{2000} - \mu_{1000}
\end{array}$$

Interpret the ANOVA results in light of this follow-up analysis.

26.28 Toxicology of lead acetate, continued. To analyze the study in Exercise 26.27, you could legitimately run a contrast test comparing the control (0 mg/kg lead acetate) to the three other treatment groups. Compute a 95% confidence interval for this contrast and conclude.

26.29 Neural basis of 3D vision. Exercise 24.32 (page 631) discussed the neural coding of stereopsis, the process of 3D vision based on the disparity of visual signals reaching both retinas. The exercise provided the response (in spikes per second)of a neuron to stereograms of 6 different disparities.

a. How many pairwise comparisons are there among the means of 6 populations?
b. Use Tukey’s method to compare these means at the overall 5% significance level.

26.30 Hooded rats: object play times. Exercise 26.10 describes an experiment to study the effects of social isolation on the behavior of hooded rats. You have analyzed the effects on social play time. Now look at another response variable, the time that a rat spends in object play during an observation period. The file *ex26-30.dat* records the time (in seconds) that each rat devoted to object play.

a. Make a plot of the 4 group means. Is there a large interaction between gender and housing type? Which main effect appears to be more important?
b. Verify that the conditions for ANOVA inference are satisfied.
c. Give the complete two-way ANOVA table. What are the *F* statistics and *P*-values for interaction and the two main effects? Explain why the test results confirm the tentative conclusions you drew from the plot of means.

26.31 Hooded rats: social play counts. The researchers who conducted the study in Exercise 26.30 also recorded the number of times each of three types of behavior (object play, locomotor play, and social play) occurred. The file *ex26-31.dat* contains the counts of social play episodes by each rat during the observation period. Use two-way ANOVA to analyze the effects of gender and housing.

26.32 Hooded rats: object play counts. The researchers who conducted the study in Exercise 26.30 also recorded the number of times each of three types of behavior (object play, locomotor play, and social play) occurred. The file *ex26-32.dat* contains the counts of object play episodes for each rat during the observation period. Carry out a complete analysis of the effects of gender and housing type.

26.33 Temperature response in goldfish. Goldfish are cold-blooded and have strong physiological responses to sudden temperature changes ("acute response"). However, they can also adapt to long-term changes in temperature over the course of weeks ("acclimation"). An experiment compared the acute response to a given test temperature (10, 12, 15, 22, or 25 degrees Celsius) in goldfish acclimated for two months to a cold (12 degrees Celsius) or warm (22 degrees Celsius) environment. Eighteen goldfish were used for each of the 10 treatment groups. Table 26.4 gives their ventilation rate in number of opercular beats per minute.[13] Here is software output for the two-way ANOVA test of these data:

```
Test temperature  F=68.13, P<0.001
Acclimation  F=12.85, P<0.001
Test temperature*Acclimation F=1.84, P=0.124
```

TABLE 26.4 Goldfish ventilation rates (number of opercular beats per minute)

Cold-acclimated fish					Warm-acclimated fish				
10 °C	**12 °C**	**15 °C**	**22 °C**	**25 °C**	**10 °C**	**12 °C**	**15 °C**	**22 °C**	**25 °C**
60	61	70	111	113	43	25	64	114	104
88	36	98	85	124	39	30	69	86	158
36	32	94	84	192	8	68	76	120	160
50	51	86	53	146	30	54	82	75	120

27	92	67	99	100	44	63	78	40	141
37	88	80	58	110	64	35	49	96	79
65	48	67	122	102	12	80	96	90	122
40	26	52	67	144	15	80	43	65	172
26	40	93	136	160	47	71	62	54	110
68	60	77	110	216	14	12	48	86	153
74	76	88	96	132	28	27	46	87	96
63	60	69	72	75	40	8	54	20	110
80	48	88	52	148	70	15	30	78	128
60	60	80	77	87	21	40	70	80	158
70	51	75	125	100	47	30	48	87	117
100	38	84	77	105	20	4	66	112	85
40	66	62	78	102	34	45	48	67	189
76	72	53	92	91	35	30	46	52	120

What are the long-term and short-term effects of temperature change on goldfish ventilation rate? Write a short conclusion based on your analysis. Follow the four-step process used in Examples 26.11 through 26.13.

26.34 Neural basis of 3D vision, continued. Exercise 26.29 discussed the neural coding of 3D vision based on image stereopsis. However, 3D vision also involves the appreciation of distances. How does the brain combine information about volume (from binocular disparity) with information about distance (from lens focus and gaze angle)? Figure 26.16 shows Minitab's interaction plot for the activity (in spikes per second) of the primary visual cortex neuron of Exercise 24.32 when visual patterns with 1 of 6 disparity values were presented on a screen located at a distance of either 20 cm, 40 cm, or 80 cm. A negative disparity corresponds to the impression of a near object. There were 12 recordings for each treatment group, collected in random order.[14] The file *ex26-34.dat* contains the raw data. Here is software output for the two-way ANOVA test of these data:

```
Distance  F=142.46, P<0.001
Disparity  F=39.79, P<0.001
Distance*Disparity interaction F=6.82, P<0.001
```

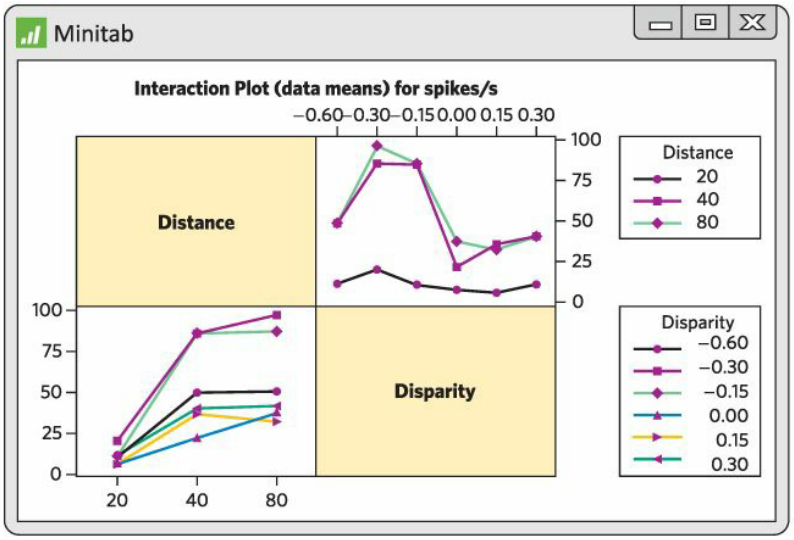


Figure 26.16
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 26.16 Plots from Minitab of the mean activity of a primary visual cortex neuron under six conditions of disparity and three visual distances, for Exercise 26.34.

How does this neuron respond to both disparity and distance cues? Write a short conclusion based on your analysis. Follow the four-step process used in Examples 26.11 through 26.13.

26.35 Temperature response in goldfish, continued. Go back to Exercise 26.33 on goldfish ventilation rate.

a. What are the degrees of freedom for the main effects and for the interaction?
b. Software gives mean squares of 39,289.4, 7411.3, and 1058.7 for test temperature, acclimation, and interaction, respectively. What is the value of MSE?
c. Compute the sums of squares. What is the value of SST? Write all your answers into a two-way ANOVA table (you can refer to Figure 26.10 for a model).
d. If you have access to software, verify that your answers match the software output (aside from rounding errors).

26.36 Neural basis of 3D vision, continued. Go back to Exercise 26.34 on neural activity.

a. What are the degrees of freedom for the main effects and for the interaction?

b. Software gives mean squares of 47,449.7, 13,253.6, and 2272.8 for distance, disparity, and interaction, respectively. What is the value of MSE?

c. Compute the sums of squares. What is the value of SST? Write all your answers into a two-way ANOVA table (you can refer to Figure 26.10 for a model).

d. The file *ex26-36.dat* contains the raw data for this experiment. If you have access to software, verify that your answers match the software output (aside from rounding errors).

26.37 Herbicide and corn hybrids. Genetic engineering has produced new corn hybrids that resist the effects of herbicides. This allows more effective control of weeds, because herbicides don't damage the corn. A study compared the effects of the herbicide glusofinate on a number of corn hybrids. The percents of necrosis (leaf burn) 10 days after application of glusofinate for several application rates (kilograms per hectare) and three corn hybrids, two resistant and one not, are provided in the file *ex26-37.dat.*[15]

a. Construct a plot of means to examine the effects of application rate and hybrid and their interaction.

b. Are the conditions for ANOVA inference satisfied? Explain.

26.38 More on dietary manipulations in fruit flies. The experiment described in Example 26.11 also examined the number of eggs produced per female. Here are the means for each of the 8 test groups:

	Amount of yeast in diet (mg)			
	0	**1**	**3**	**7**
Wild-type	6.4	15.1	33.8	56.3
Mutant	7.7	20.8	50.8	78.3

a. Plot the means on a single graph and describe the main effects and interaction.

b. A two-way ANOVA analysis gives significant *P*-values for both main effects and the interaction effect. However, although there are no outliers and no major deviation from Normality in the data, the 8 standard deviations range from 2.1 to 12.0 eggs per female. What do you make of this information? How might it affect your interpretation of the ANOVA results?

26.39 Comparisons among means for one factor in a two-way analysis. We have illustrated the Tukey pairwise comparisons among all treatment means in both one-way and two-way settings. The method can also compare the mean responses to just one of the two factors in a two-way setting. Just do a one-way ANOVA on the two-way data with only one factor listed. This combines data for all levels of the other factor, so it is useful only when interactions are small. Return to the data on phosphorus in tomatoes, Table 26.2. Do a one-way ANOVA that uses all 36 observations, with fertilizer type (nitrogen level) as the only factor. Ask for Tukey pairwise comparisons among the three levels of nitrogen, with overall confidence level 95%.

a. How many observations per group does your analysis use?

b. What do you conclude from the *F* statistic and its *P*-value?

c. Which pairwise differences of means for the 3 nitrogen levels are significant at the overall 5% level?

d. Do you think these pairwise comparisons are useful for these data? (*Hint:* What population does each of the three samples represent?)

CHAPTER 27 Nonparametric Tests

Blickwinkel/Alamy

IN THIS CHAPTER WE COVER...

- Comparing two samples: the Wilcoxon rank sum test
- Matched pairs: the Wilcoxon signed rank test
- Comparing several samples: the Kruskal-Wallis test

The most commonly used methods of inference about the means of quantitative response variables assume that the variables in question have Normal distributions in the population or populations from which we draw our data. In practice, of course, no distribution is exactly Normal. Fortunately, our usual methods of inference about population means (the one-sample and two-sample *t* procedures and analysis of variance) are quite robust. That is, the results of inference are not very sensitive to moderate lack of Normality, especially when the samples are reasonably large. Practical guidelines for taking advantage of the robustness of these methods appear in Chapters 17, 18, and 24.

What can we do if plots suggest that the data are clearly not Normal, especially when we have only a few observations? This is not a simple question. Here are the basic options:

1. If lack of Normality is due to outliers, it may be legitimate to remove outliers if you have reason to think that they do not come from the same population as the other observations. Equipment failure that produced a bad measurement, for example, entitles you to remove the outlier and analyze the remaining data. *But if an outlier appears to be "real data," you should not arbitrarily remove it.* The discussion topic in Chapter 2 (page 55) addresses in more depth how to recognize and treat outliers.

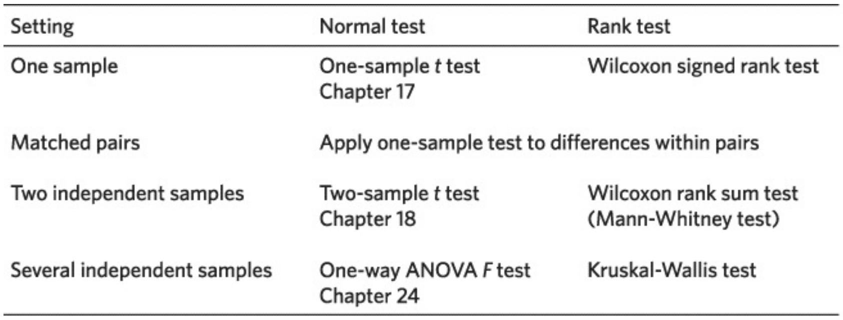
CAUTION

2. Sometimes we can transform our data so that their distribution is more nearly Normal. Transformations such as the logarithm that shorten the long tail of right-skewed distributions are particularly helpful. We used thelogarithm transformation in Example 4.2 (page 96).
3. In some settings, other standard distributions replace the Normal distributions as models for the overall pattern in the population. The lifetimes in service of equipment and the survival times of cancer patients after treatment usually have right-skewed distributions. Statistical studies in these areas use families of right-skewed distributions rather than Normal distributions. There are inference procedures for the parameters of these distributions that replace the t procedures.
4. Modern bootstrap methods and permutation tests use heavy computing to avoid requiring Normality or any other specific form of sampling distribution. We recommend these methods unless the sample is so small that it may not represent the population well. For an introduction, see the Companion Chapter 16 of the somewhat more advanced text *Introduction to the Practice of Statistics,* available online at www.macmillanlearning.com/ips9e.
5. Finally, there are other nonparametric methods, which do not assume anyspecific form for the distribution of the population. Unlike bootstrap andpermutation methods, common nonparametric methods do not make

useof the actual values of the observations.

This chapter concerns one type of nonparametric procedure: tests that can replace the t tests and one-way analysis of variance *when the Normality conditions for those tests are either not met or questionable.* The most useful nonparametric tests are rank tests based on the rank (place in order) of each observation in the set of all the data.

Figure 27.1 presents an outline of the standard tests (based on Normal distributions) and the rank tests that compete with them. These rank tests requirethat the population or populations have *continuous distributions.* That is, each distribution must be described by a *density curve* (Chapter 9, page 225) that allows observations to take any value within some interval of outcomes. The Normal curves are one shape of density curve. Rank tests allow curves of any shape.



Setting	Normal test	Rank test
One sample	One-sample t test Chapter 17	Wilcoxon signed rank test
Matched pairs	Apply one-sample test to differences within pairs	
Two independent samples	Two-sample t test Chapter 18	Wilcoxon rank sum test (Mann-Whitney test)
Several independent samples	One-way ANOVA F test Chapter 24	Kruskal-Wallis test

Figure 27.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.1 Comparison of tests based on Normal distributions with rank tests for similar settings.

The rank tests we will study concern the *center* of a population or populations. When a population has at least roughly a Normal distribution, we describe its center by the mean. The "Normal tests" in Figure 27.1 all test hypotheses about population means. When distributions are strongly skewed, we often prefer the median to the mean as a measure of center. In their simplest form, the hypotheses for rank tests just replace mean with median.

Webegin by describing the most common rank test, the one for comparing two samples. In this setting we also explain ideas common to all rank tests: the big idea of using ranks, the conditions required by rank tests, the nature of the hypotheses tested, and the difference between exact distributions for use with small samples and Normal approximations for use with larger samples.

COMPARING TWO SAMPLES: THE WILCOXON RANK SUM TEST

Two-sample problems (see Chapter 18) are among the most common in statistics. The most useful nonparametric significance test compares two distributions. Here is an example of this setting.

EXAMPLE 27.1 Weeds among the corn

STATE: Does the presence of small numbers of weeds reduce the yield of corn? Lamb’s-quarter is a common weed in corn fields. A researcher planted corn at the same rate in 8 small plots of ground, then weeded the corn rows by hand to allow no weeds in 4 randomly selected plots and exactly 3 lamb’s-quarter plants per meter of row in the other 4 plots. Here are the yields of corn (bushels per acre) in each of the plots:[1]

0 weeds per meter	166.7	172.2	165.0	176.9
3 weeds per meter	158.6	176.4	153.1	156.0

PLAN: Make a graph to compare the two sets of yields. Test the null hypothesis that there is no difference against the one-sided alternative that yields are higher when no weeds are present.

SOLVE (first steps): A dotplot (Figure 27.2) suggests that yields may be higher when there are no weeds. There is one outlier; because it is correct data, we cannot remove it. The samples are too small to rely on the robustness of the two-sample *t* test. We will now develop a test that does not require Normality in order to be valid.

Blickwinkel/Alamy

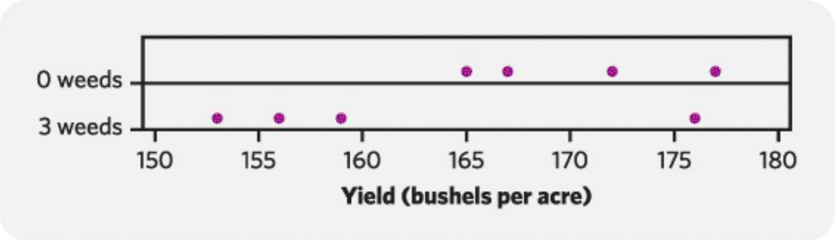


Figure 27.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.2 Dotplot of corn yields from plots with no weeds and with 3 weeds per meter of row.

First, rank all 8 observations together. To do this, arrange them in order from smallest to largest:

153.1 156.0 158.6 **165.0** **166.7** **172.2** 176.4 **176.9**

The boldface entries in the list are the yields with no weeds present. We see that four of the five highest yields come from that group, suggesting that yields are higher with no weeds. The idea of rank tests is to look just at position in this ordered list. To do this, replace each observation with its order, from 1 (smallest) to 8 (largest). These numbers are the *ranks:*

Yield	153.1	156.0	158.6	**165.0**	**166.7**	**172.2**	176.4	**176.9**
Rank	1	2	3	**4**	**5**	**6**	7	**8**

RANKS

To rank observations, first arrange them in order from smallest to largest. The **rank** of each observation is its position in this ordered list, starting with rank 1 for the smallest observation.

Moving from the original observations to their ranks retains only the ordering of the observations and makes no other use of their numerical values. Working with ranks allows us to dispense with specific conditions on the shape of the distribution, such as Normality.

If the presence of weeds reduces corn yields, we expect the ranks of the yields from plots without weeds to be larger as a group than the ranks from plots with weeds. Let’s compare the *sums* of the ranks from the two treatments:

Treatment	**Sum of ranks**
No weeds	23
Weeds	13

These sums measure how much the ranks of the weed-free plots as a group exceed those of the weedy plots. In fact, the sum of the ranks from 1 to 8 is equal to 36, so it is enough to report the sum for one of the two groups. If the sum of the ranks for the weed-free group is 23, the ranks for the other group must add to 13, because $23+13=36$. If the weeds had no effect, we would expect the sum of the ranks in each group to be 18 (half of 36). Here are the facts we need in a more general form that takes account of the fact that our two samples need not be the same size.

THE WILCOXON RANK SUM TEST

Draw an SRS of size n_1 from one population and draw an independent SRS of size n_2 from a second population. There are N observations in all, where $N = n_1 + n_2$. Rank all N observations. The sum W of the ranks for the first sample is the **Wilcoxon rank sum statistic.** If the two populations have the same continuous distribution, then W has mean

$$\mu_W = \frac{n_1(N+1)}{2}$$

$$\mu_W = \frac{n_1 (N+1)}{2}$$

and standard deviation

σW=n1n2(N+1)12

$$\sigma_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

The **Wilcoxon rank sum test** rejects the hypothesis that the two populations have identical distributions when the rank sum *W* is far from its mean.

In the corn yield study of Example 27.1, we want to test the hypotheses

H_0: no difference in distribution of yields
H_a: yields are systematically higher in weed-free plots

Our test statistic is the rank sum W=23 $W = 23$ for the weed-free plots.

EXAMPLE 27.2 Weeds among the corn: inference

SOLVE: First, note that the conditions for the Wilcoxon test are met: The data come from a randomized comparative experiment, and the yield of corn in bushels per acre has a continuous distribution.

There are N=8 $N = 8$ observations in all, with n1=4 $n_1 = 4$ and n2=4 $n_2 = 4$. The sum of ranks for the weed-free plots has mean

μW=n1(N+1)2=(4)(9)2=18

$$\begin{aligned}\mu_W &= \frac{n_1 (N+1)}{2} \\ &= \frac{(4)(9)}{2} = 18\end{aligned}$$

and standard deviation

σW=n1n2(N+1)12=(4)(4)(9)12=12=3.464

$$\begin{aligned}\sigma_W &= \sqrt{\frac{n_1 n_2 (N+1)}{12}} \\ &= \sqrt{\frac{(4)(4)(9)}{12}} = \sqrt{12} = 3.464\end{aligned}$$

Although the observed rank sum W=23 $W = 23$ is higher than the mean, it is only about 1.4 standard deviations higher. We now suspect that the data do not give strong evidence that yields are higher in the population of weed-free corn.

The *P*-value for our one-sided alternative is P(W≥23)$P(W \geq 23)$, the probability that *W* is at least as large as the value for our data when H_0 is true. Software tells us that this probability is P=0.1$P = 0.1$.

CONCLUDE: The data provide some evidence (P=0.1$P = 0.1$) that corn yields are lower when weeds are present. There are only 4 observations in each group, so even quite large effects can fail to reach the levels of significance usually considered convincing, such as P<0.05$P < 0.05$. A larger experiment might clarify the effect of weeds on corn yield.

APPLY YOUR KNOWLEDGE

27.1 Daily activity and obesity. Our lead example for the two-sample *t* procedures in Chapter 18 concerned a study comparing the level of physical activity of lean and mildly obese people who don't exercise. Here are the minutes per day that the subjects spent standing or walking over a 10-day period:

Lean subjects		Obese subjects	
511.100	543.388	260.244	416.531
607.925	677.188	464.756	358.650
319.212	555.656	367.138	267.344
584.644	374.831	413.667	410.631
578.869	504.700	347.375	426.356

The data are a bit irregular but not distinctly non-Normal. Let's use the Wilcoxon test for comparison with the two-sample *t* test.

a. Find the median minutes spent standing or walking for each group. Which group appears more active?
b. Arrange all 20 observations in order and find the ranks.
c. Take *W* to be the sum of the ranks for the lean group. What is the

value of W? If the null hypothesis (no difference between the groups) is true, what are the mean and standard deviation of W?

d. Does comparing W with the mean and standard deviation suggest that the lean subjects are more active than the obese subjects?

27.2 Immune response. NOD receptors are involved in the immune response to bacterial infections. Researchers investigated the impact of a deleterious mutation (Bid$^{-/-}$) on NOD-mediated immune signaling in mice. Here are the immune responses to a NOD trigger in wild-type and Bid$^{-/-}$ mice, as assessed by the serum concentration of a specific protein marker (in picograms per milliliter, pg/ml):[2]

Wild-type	17.3	22.8	30.8	38.0	55.1
Bid$^{-/-}$	45.9	27.1	25.7	27.7	38.8

a. The sample sizes are small. We may prefer a nonparametric test to compare the groups. Find the Wilcoxon rank sum W for the wild-type group, along with its mean and standard deviation under the null hypothesis (no difference between groups).

b. Do you think that W is far enough from the mean under H_0 to suggest that there may be a difference between the groups?

The Normal approximation for WW To calculate the P-value $P(W \geq 23)$ $P(W \geq 23)$ for Example 27.2, we need to know the sampling distribution of the rank sum W when the null hypothesis is true. This distribution depends on the two sample sizes n_1 and n_2. Tables of such values are therefore unwieldy, though you can find them in handbooks of statistical tables. Most statistical software will give you P-values, as well as carry out the ranking and calculate W. However, many software packages give only approximate P-values. You must learn what your software offers.

With or without software, P-values for the Wilcoxon test are often based on the fact that **the rank sum statistic W becomes approximately Normal as the two sample sizes increase.** We can then form yet another z statistic by standardizing W:

$$z=W-\mu W\sigma W=W-n1(N+1)/2n1n2(N+1)/12$$

$$\begin{aligned} z &= \frac{W-\mu_W}{\sigma_W} \\ &= \frac{W-n_1(N+1)/2}{\sqrt{n_1 n_2(N+1)/12}} \end{aligned}$$

Use standard Normal probability calculations to find *P*-values for this statistic. Because *W* takes only whole-number values, an idea called the *continuity correction* improves the accuracy of the approximation.

CONTINUITY CORRECTION

To apply the **continuity correction** in a Normal approximation for a variable that takes only whole-number values, act as if each whole number occupies the entire interval from 0.5 below the number to 0.5 above it.

EXAMPLE 27.3 Weeds among the corn: Normal approximation

The standardized rank sum statistic *W* in our corn yield example is

z=W-μWσW=23-183.464=1.44

$$z = \frac{W-\mu_W}{\sigma_W} = \frac{23-18}{3.464} = 1.44$$

We expect *W* to be larger when the alternative hypothesis is true, so the approximate *P*-value is (from Table B)

P(Z≥1.44)=0.0749

$$P(Z \geq 1.44) = 0.0749$$

We can improve this approximation by using the continuity correction. To do this, act as if the whole number 23 occupies the entire interval from 22.5 to 23.5. Calculate the *P*-value P(W≥23)$P(W \geq 23)$ as P(W≥22.5)$P(W \geq 22.5)$, because the value 23 is included in the range whose probability we want. Here is the calculation:

P(W≥22.5)=PW-μWσW≥22.5-183.464=P(Z≥1.30)=0.0968

$$\begin{aligned} P(W \geq 22.5) &= P\left(\frac{W-\mu_W}{\sigma_W} \geq \frac{22.5-18}{3.464}\right) \\ &= P(Z \geq 1.30) \\ &= 0.0968 \end{aligned}$$

> using Table B (we get $P = 0.0970$ using technology for a more accurate Normal computation). If you do not use the exact distribution of W (from software or tables), you should always use the continuity correction in calculating P-values.

APPLY YOUR KNOWLEDGE

27.3 Daily activity and obesity, continued. In Exercise 27.1 you found the Wilcoxon rank sum W and its mean and standard deviation. We want to test the null hypothesis that the two groups don't differ in activity against the alternative hypothesis that the lean subjects spend more time standing and walking.

a. What is the probability expression for the P-value of W if we use the continuity correction?
b. Find the P-value. What do you conclude?

27.4 Immune response, continued. Use your values of W, μ_W, and σ_W from Exercise 27.2 to see whether the Bid$^{-/-}$ mutation has an effect on immune response.

a. The two-sided P-value is $2P(W \leq ?)$. Using the continuity correction, what number replaces the ? in this probability?
b. Find the P-value. What do you conclude about the effect of the Bid$^{-/-}$ mutation?

27.5 Fruit fly reproduction. Reproduction has a high physiological cost. Researchers hypothesized that there is a trade-off between longevity and reproduction. They designed an experiment comparing the reproductive output of wild-type fruit flies and mutants with a longer reproductive cycle and higher longevity. Here are the number of eggs produced per female when the fruit flies were fed a protein-rich diet:[3]

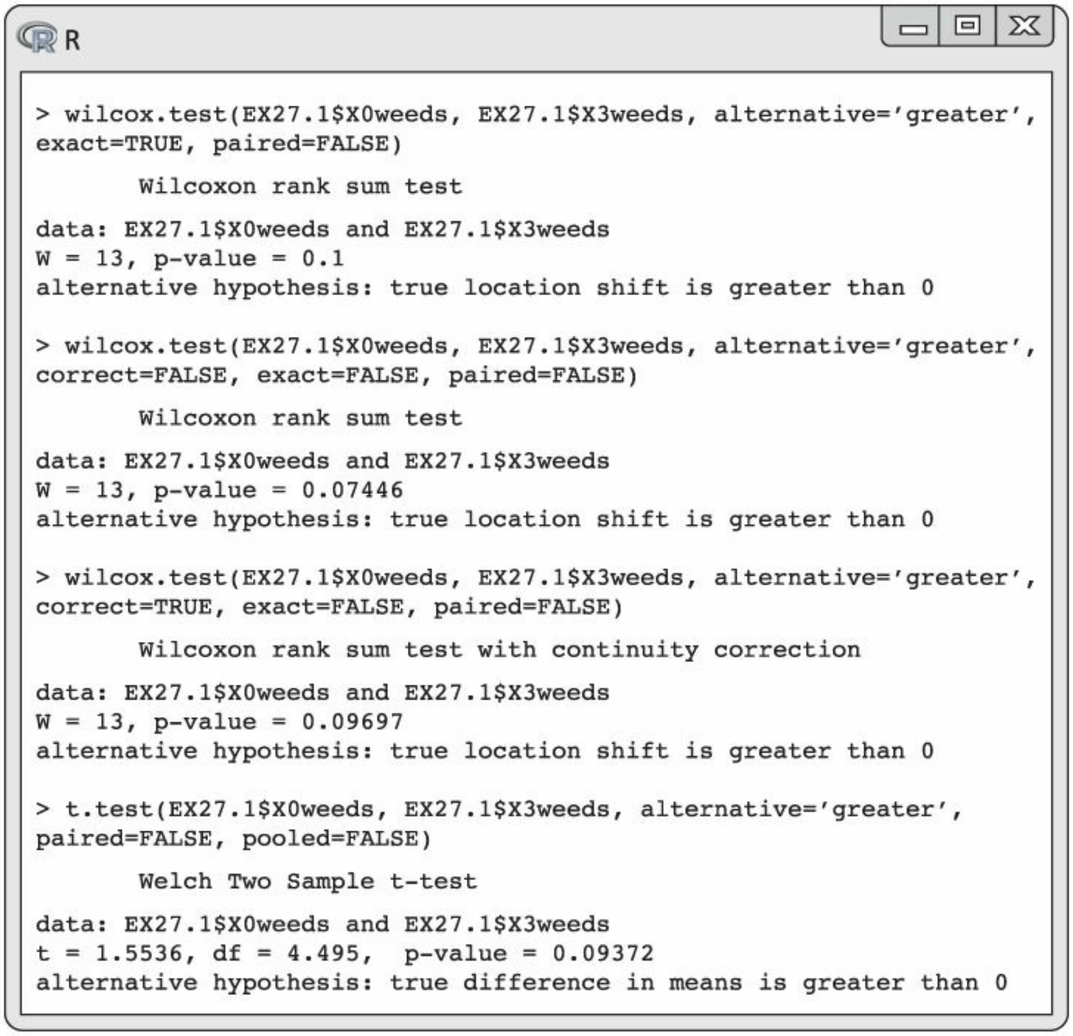
Wild-type	63.80	86.45	82.60	73.35	85.50
Mutant	40.35	67.95	48.50	64.25	60.35

Do the data provide evidence that the mutants invest less in reproduction (as measured by the number of eggs produced per female) than the wild-type fruit flies? Follow the four-step process as illustrated in Examples 27.1 and 27.2.

Using technology For samples as small as those in the corn yield study of Example 27.1, we prefer software that gives the exact *P*-value for the Wilcoxon test rather than the Normal approximation. Neither the Excel spreadsheet nor the TI-83 calculator has menu entries for rank tests. Some statistical programs offer only the Normal approximation for the Wilcoxon test. Other statistical programs carry out the Mann-Whitney test rather than the Wilcoxon test. The two tests always have the same *P*-value, because the two test statistics are related by simple algebra.

EXAMPLE 27.4 Weeds among the corn: software output

Figure 27.3 displays output from the software program R for the corn yield data. The top three queries reflect the three methods available to obtain the *P*-value of the Wilcoxon sum rank test: exact, Normal, and Normal with a continuity correction, respectively. The exact *P*-value is P=0.1$P = 0.1$. The Normal approximation with continuity correction, P=0.097$P = 0.097$ in Example 27.3 and in the third R query, is quite accurate.



```
> wilcox.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
exact=TRUE, paired=FALSE)

        Wilcoxon rank sum test

data: EX27.1$X0weeds and EX27.1$X3weeds
W = 13, p-value = 0.1
alternative hypothesis: true location shift is greater than 0

> wilcox.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
correct=FALSE, exact=FALSE, paired=FALSE)

        Wilcoxon rank sum test

data: EX27.1$X0weeds and EX27.1$X3weeds
W = 13, p-value = 0.07446
alternative hypothesis: true location shift is greater than 0

> wilcox.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
correct=TRUE, exact=FALSE, paired=FALSE)

        Wilcoxon rank sum test with continuity correction

data: EX27.1$X0weeds and EX27.1$X3weeds
W = 13, p-value = 0.09697
alternative hypothesis: true location shift is greater than 0

> t.test(EX27.1$X0weeds, EX27.1$X3weeds, alternative='greater',
paired=FALSE, pooled=FALSE)

        Welch Two Sample t-test

data: EX27.1$X0weeds and EX27.1$X3weeds
t = 1.5536, df = 4.495,  p-value = 0.09372
alternative hypothesis: true difference in means is greater than 0
```

Figure 27.3

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.3 Output from *R* for the data in Example 27.1. The output compares the three methods for computing the *P*-value of the Wilcoxon rank sum test and the *P*-value obtained with a two-sample *t* test.

The last query in Figure 27.3 is the two-sample *t* test from Chapter 18, which does not assume that the two populations have the same standard deviation. It gives P=0.0937$P = 0.0937$, close to the Wilcoxon *P*-value. Because the *t* test is quite robust, it is somewhat unusual for *P*-values derived from *t* and those from *W* to differ greatly.

APPLY YOUR KNOWLEDGE

27.6 Immune response: software. Use your software to repeat the Wilcoxon test you did in Exercise 27.4. By comparing the results, state how your software finds *P*-values for *W*: exact distribution, Normal approximation with continuity correction, or Normal approximation without continuity correction.

27.7 Daily activity and obesity: software. Use your software to carry out the one-sided Wilcoxon rank sum test that you did by hand in Exercise 27.3. Use the exact distribution if your software will do it. Compare the software result with your result in Exercise 27.3.

27.8 Weeds among the corn. The corn yield study of Example 27.1 also examined yields in four plots having 9 lamb's-quarter plants per meter of row. The yields (bushels per acre) in these plots were

162.8 142.4 162.7 162.4

There is a clear outlier, but rechecking the results found that this is the correct yield for this plot. The outlier makes us hesitant to use *t* procedures because x-$\bar{x}$ and *s* are not resistant to outliers.

a. Is there evidence that the presence of 9 weeds per meter reduces corn yields when compared with weed-free corn? Use the Wilcoxon rank sum test with the data above and part of the data from Example 27.1 to answer this question.
b. Compare the results from a with those from the two-sample *t* test for these data.
c. Now remove the low outlier 142.4 from the data with 9 weeds per meter. Repeat both the Wilcoxon and *t* analyses. By how much did the outlier reduce the mean yield in its group? By how much did it increase the standard deviation? Did it have a practically important impact on your conclusions?

What hypotheses does Wilcoxon test? Our null hypothesis is that weeds do not affect yield. The alternative hypothesis is that yields are lower when weeds are present. If we are willing to assume that yields are Normally distributed, or if we have reasonably large samples, we can use the two-sample *t* test for means. Our hypotheses then have the form

H0:μ1=μ2Ha:μ1>μ2

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

When the distributions may not be Normal, we might restate the hypotheses in terms of population medians rather than means:

H0: population median1=population median2Ha: population median1>population median2

$$H_0 : \text{population median}_1 = \text{population median}_2$$
$$H_a : \text{population median}_1 > \text{population median}_2$$

The Wilcoxon rank sum test provides a test of these median-based hypotheses, but only if an additional condition is met: Both populations must have distributions of *the same shape.* That is, the density curve for corn yields with 3 weeds per meter looks exactly like that for yields with no weeds except that it may slide to a different location on the scale of yields.

The same-shape condition is too strict to be reasonable in practice. Fortunately, the Wilcoxon test also applies in a more useful setting. It compares *any two* continuous distributions, whether or not they have the same shape, by testing hypotheses that we can state in words as

H0: the two distributions are the sameHa: one has values that are systematically larger

$$H_0 : \text{the two distributions are the same}$$
$$H_a : \text{one has values that are systematically larger}$$

A more exact statement of the "systematically larger" alternative hypothesis is a bit tricky, so we won't try to give it here.[4] The R output of Figure 27.3 states the hypotheses in terms of location shift. These hypotheses really are "nonparametric," because they do not involve any specific parameter such as the mean or median. If the two distributions do have the same shape, the general hypotheses are reduced to comparing medians. *Many texts and computer outputs state the hypotheses in terms of medians, sometimes ignoring the same-shape condition.* We recommend that you express the hypotheses in words rather than symbols. "Yields are systematically higher in weed-free plots" is easy tounderstand and is a good statement of the

CAUTION

effect that the Wilcoxon test looks for.

Why don't we discuss the confidence intervals for the difference in population medians that some statistical programs offer? These intervals require the unrealistic same-shape condition. The more general "systematically larger" hypothesis does not involve a specific parameter, so there is no accompanying confidence interval.

APPLY YOUR KNOWLEDGE

27.9 Daily activity and obesity: hypotheses. We could use either two-sample t or the Wilcoxon rank sum to test the null hypothesis that lean and mildly obese people don't differ in the time they spend standing and walking against the alternative hypothesis that lean people generally spend more time in these activities. Explain carefully what H_0 and H_a are for t and for W.

27.10 Immune response: hypotheses. We are interested in whether the deleterious mutation Bid$^{-/-}$ changes the serum concentration of a protein marker of NOD-mediated immune signaling in mice "on average."

a. State null and alternative hypotheses in terms of population means. What test would we typically use for these hypotheses? What conditions does this test require?

b. State null and alternative hypotheses in terms of population medians. What test would we typically use for these hypotheses? What conditions does this test require?

Dealing with ties in rank tests We have chosen our examples and exercises to this point rather carefully: They all involve data in which *no two values are the same.* This allowed us to rank all the values. In practice, however, we often find observations tied at the same value. What shall we do? The usual practice is to *assign all tied values the* average *of the ranks they occupy*. Here is an example with 6 observations:

Observation	153	155	158	158	161	164

Rank	1	2	3.5	3.5	5	6

The tied observations occupy the third and fourth places in the ordered list, so they share rank 3.5.

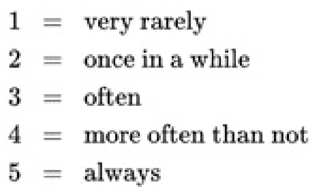
The exact distribution for the Wilcoxon rank sum W applies only to data without ties. Moreover, the standard deviation σ_W must be adjusted if ties are present. The Normal approximation can be used after the standard deviation is adjusted. Statistical software will detect ties, make the necessary adjustment, and switch to the Normal approximation. *In practice, software is required to use rank tests when the data contain tied values.*

Some data have many ties because the scale of measurement has only a few values. Rank tests are often used for such data. Here is an example.

EXAMPLE 27.5 Food safety at fairs

STATE: Food sold at outdoor fairs and festivals may be less safe than food sold in restaurants because it is prepared in temporary locations and often by volunteer help. What do people who attend fairs think about the safety of the food served? One study asked this question of people at a number of fairs in the Midwest: “How often do you think people become sick because of food they consume prepared at outdoor fairs and festivals?” The possible responses were

1=very rarely2=once in a while3=often4=more often than not5=always



1 = very rarely
2 = once in a while
3 = often
4 = more often than not
5 = always

In all, 303 people answered the question. Of these, 196 were women and 107 were men.[5] We suspect that women are more concerned than men about food safety. Is there good evidence for this conclusion?

PLAN: Do data analysis to understand the difference between women and men. Because the 1-to-5 ratings are in fact rankings rather than actual quantitative data, we should use a statistical procedure involving ranks. Check the conditions required by the Wilcoxon test. If the conditions are met, use the Wilcoxon test for the hypotheses

H0: men and women do not differ in their responsesHa: women give systematically higher responses than men

H_0 : men and women do not differ in their responses
H_a : women give systematically higher responses than men

SOLVE: Here are the data, presented as a two-way table of counts:

	Response					**Total**
	1	**2**	**3**	**4**	**5**	
Female	13	108	50	23	2	196
Male	22	57	22	5	1	107
Total	35	165	72	28	3	303

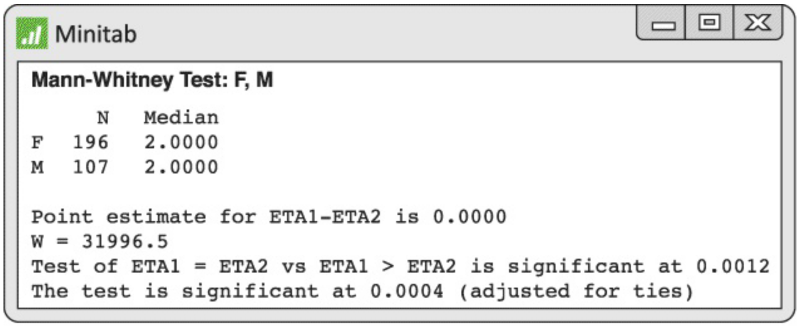
Comparing row percents shows that the women in the sample do tend to give higher responses (showing more concern):

	Response					**Total**
	1	**2**	**3**	**4**	**5**	
Percent of females	6.6	55.1	25.5	11.7	1.0	100
Percent of males	20.6	53.3	20.6	4.7	1.0	100

Are these differences between women and men statistically significant?

The most important condition for inference is that the subjects be a *random sample* of people who attend fairs, at least in the Midwest. The researcher visited 11 different fairs. She stood near the entrance and stopped every 25th adult who passed. Because no personal choice was involved in choosing the subjects, we can reasonably treat the data as coming from a random sample. (As usual, there was some nonresponse, which could create bias.) The Wilcoxon test also requires that responses have *continuous distributions*. We think that the subjects' opinions about how often people become sick from food at fairs really do form a continuous distribution. The questionnaire asks them to round off their opinions to the nearest value in the five-point scale. So we are willing to use the Wilcoxon test.

Because the responses can take only five values, there are many ties. All 35 people who chose "very rarely" are tied at 1, and all 165 who chose “once in a while” are tied at 2. Figure 27.4 gives output from Minitab. The Wilcoxon (reported as Mann-Whitney) test for the one-sided alternative that women are more concerned about food safety at fairs is highly significant ($P = 0.0004$, adjusted for tie).



Mann-Whitney Test: F, M

```
       N   Median
F    196   2.0000
M    107   2.0000

Point estimate for ETA1-ETA2 is 0.0000
W = 31996.5
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0.0012
The test is significant at 0.0004 (adjusted for ties)
```

Figure 27.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.4 Output from Minitab for the data of Example 27.5.

CONCLUDE: There is very strong evidence (P=0.0004$P = 0.0004$) that women are more concerned than men about the safety of food served at fairs.

Because the sample sizes are so large in Example 27.5, a *t* test would have given similar results. However, the *t* statistic treats the response values 1 through 5 as meaningful numbers. In particular, the possible responses are treated as though they are equally spaced. The difference between “very rarely” and “once in a while” is considered the same as the difference between “once in a while” and “often.” This may not make sense. The rank test, on the other hand, uses only the order of the responses, not their actual values. The responses are arranged in order from least to most concerned about safety, so the rank test makes sense. *Statisticians often avoid using t procedures when there is not a fully meaningful scale of measurement.*

CAUTION

Because we have a two-way table, we might have applied the chi-square test (Chapter 22), which asks if there is a significant relationship *of any kind* between gender and response. The chi-square test ignores the ordering of the responses and so doesn’t tell us whether women are *more* concerned than men about the safety of the food served. This question depends on the ordering of responses from least concerned to most concerned.

APPLY YOUR KNOWLEDGE

Software is required to adequately carry out the Wilcoxon rank sum test in the presence of ties. All of the following exercises concern data with ties.

27.11 Do birds learn to time their breeding? Blue titmice eat caterpillars but do they adjust their breeding based on the peak caterpillar supply in the previous year? Researchers randomly assigned 7 bird pairs to have the natural caterpillar supply supplemented when the birds were feeding their young and another 6 pairs to serve as a control group relying only on the natural food supply. The next year, they measured how many days after the caterpillar peak the birds produced their nestlings.[6] Here are the data (days after the caterpillar peak):

Control	4.6	2.3	7.7	6.0	4.6	−1.2	
Supplemented	15.5	11.3	5.4	16.5	11.3	11.4	7.7

The null hypothesis is no difference in timing; the alternative hypothesis is that the supplemented birds miss the peak by more days because they don’t adjust their breeding date.

a. There are three sets of ties, at 4.6, 7.7, and 11.3. Arrange the observations in order and assign average ranks to each tied observation.

b. Take W to be the rank sum for the supplemented group. What is the value of W?

c. Use software to find the P-value of the Wilcoxon test, and conclude.

Hugh Clark/Frank Lane Picture Agency /CORBIS

27.12 Mycorrhizal colonies and plant nutrition. Mycorrhizal fungi are present in the roots of many plants. In this symbiotic relationship, the plant supplies nutrition to the fungus and the fungus helps the plant absorb nutrients from the soil. An experiment compared two genetic types of tomato plants, a wild-type variety that is susceptible to mycorrhizal colonies and a mutant that is not. In particular, researchers expect that, in the absence of fertilizer, plants with mycorrhizal fungi would better be able to absorb what little nitrogen might be available in the soil. Here are plant nitrogen contents at harvest time (in percent of whole dry mass):[7]

Wild-type	1.21	1.57	1.30	1.19	1.23	1.29
Mutant	1.23	1.05	1.28	1.04	1.25	1.21

a. Arrange the observations in order and find their ranks.

b. Take *W* to be the rank sum for wild-type tomato plants. What is the value of *W*?

c. Use software: Does *W* provide significant evidence that the nitrogen contents of the wild-type tomato plants with mycorrhizal fungi are systematically larger than those in mutants without mycorrhizal fungi?

27.13 Good smells and purchasing behavior. Many aspects of human behavior are influenced by our environment. One study of this phenomenon took place in a small pizzeria in France on two Saturday evenings in May. On one of these evenings, a relaxing lavender odor was spread through the restaurant. The two evenings were comparable in many ways (weather, customer count, and so on), so we are willing to regard the data as independent SRSs from spring Saturday evenings at this restaurant. The authors say, “Therefore at this stage it would be impossible to generalize the results to other restaurants.” Here are the amounts spent (in euros) by customers:[8]

No odor									
15.9	18.5	15.9	18.5	18.5	21.9	15.9	15.9	15.9	15.9
15.9	18.5	18.5	18.5	20.5	18.5	18.5	15.9	15.9	15.9
18.5	18.5	15.9	18.5	15.9	18.5	15.9	25.5	12.9	15.9

Lavender odor									
21.9	18.5	22.3	21.9	18.5	24.9	18.5	22.5	21.5	21.9
21.5	18.5	25.5	18.5	18.5	21.9	18.5	18.5	24.9	21.9
25.9	21.9	18.5	18.5	22.8	18.5	21.9	20.7	21.9	22.5

Examine the data and comment on departures from Normality. Is there significant evidence that the lavender odor encourages customers to spend more? Follow the four-step process.

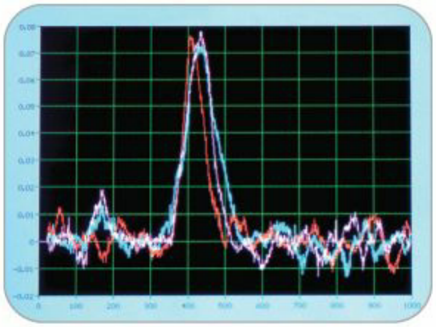
MATCHED PAIRS: THE WILCOXON SIGNED RANK TEST

We use the one-sample *t* procedures for inference about the mean of one population or for inference about the mean difference in a matched pairs setting. The matched pairs setting is more important, because good studies are generally comparative. We will now look at a rank test for this setting.

EXAMPLE 27.6 Visual receptive fields

STATE: Neurons in the visual cortex fire an increased number of action potentials (“spikes”) when a specific area of the retina is visually stimulated. This area is called the neuron’s receptive field. A neuron in the primary visual cortex is recorded while a monkey first looks at a dot on a screen (spontaneous activity, SA) and then looks at a pattern of random dots (response, R). Here is the neural activity in number of spikes per second for 9 different recordings of that neuron:[9]

Recording	1	2	3	4	5	6	7	8	9
SA	2.5	7.5	10.0	0.0	12.5	2.5	0.0	2.5	17.5
R	16.7	20.0	23.3	16.7	56.7	0.0	26.7	36.7	10.0
Difference (R − SA)	14.2	12.5	13.3	16.7	44.2	−2.5	26.7	34.2	−7.5



Donald Pye/Alamy

The neuron’s receptive field is activated if the neuron’s response is systematically higher following presentation of the visual pattern (R) than just before (SA). Is this neuron’s receptive field activated by the visual pattern presented?

PLAN: We would like to test the hypotheses

H0: neural activity has the same distribution for both SA and RHa: neural activity is systematically higher for R

H_0 : neural activity has the same distribution for both SA and R
H_a : neural activity is systematically higher for R

SOLVE (first steps): Because this is a matched pairs design, we base our inference on the differences. The matched pairs *t* test gives t=3.085 $t = 3.085$ with one-sided *P*-value P=0.007$P = 0.007$. We cannot reliably assess Normality from so few observations. We would therefore like to use a rank test.

Positive differences in Example 27.6 indicate that the neuron’s activity was stronger during the stimulus presentation (R). If neural activity is generally higher for the response to stimulus, the positive differences should be farther from zero in the positive direction than the negative differences are in the negative direction. We therefore compare the absolute values of the differences, that is, their magnitudes without a sign. Here they are, with boldface indicating the positive values:

14.2 **12.5** **13.3** **16.7** **44.2** 2.5 **26.7** **34.2** 7.5

Arrange these in increasing order and assign ranks, keeping track of which values were originally positive. Tied values receive the average of their ranks. If there are zero differences, discard them before ranking.

Absolute value	2.5	7.5	**12.5**	**13.3**	**14.2**	**16.7**	**26.7**	**34.2**	**44.2**
Rank	1	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**

The test statistic is the sum of the ranks of the positive differences. (We could equally well use the sum of the ranks of the negative differences.) This is the *Wilcoxon signed rank statistic*. Its value here is W+=42$W^+ = 42$.

THE WILCOXON SIGNED RANK TEST FOR MATCHED PAIRS

Draw an SRS of size *n* from a population for a matched pairs study and take the differences in responses within pairs. Rank the absolute values of these differences. The sum W^+ of the ranks for the positive differences is the **Wilcoxon signed rank statistic.** If the distribution of the responses is not affected by the different treatments within pairs, then W^+ has mean

μW+=n(n+1)4

$$\mu_{W^+} = \frac{n(n+1)}{4}$$

and standard deviation

σW+=n(n+1)(2n+1)24

$$\sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

The **Wilcoxon signed rank test** rejects the hypothesis that there are no systematic differences within pairs when the rank sum W^+ is far from its mean.

EXAMPLE 27.7 Visual receptive fields, continued

SOLVE: In the neural receptive field study of Example 27.6, $n = 9$. If the null hypothesis (no systematic effect of the visual stimulus) is true, the mean of the signed rank statistic is

μW+=n(n+1)4=(9)(10)4=22.5

$$\mu_{W^+} = \frac{n(n+1)}{4} = \frac{(9)(10)}{4} = 22.5$$

The standard deviation of W^+ under the null hypothesis is

σW+=n(n+1)(2n+1)24=(9)(10)(19)24=71.25=8.441

$$\begin{aligned}\sigma_{W^+} &= \sqrt{\frac{n(n+1)(2n+1)}{24}}\\ &= \sqrt{\frac{(9)(10)(19)}{24}}\\ &= \sqrt{71.25} = 8.441\end{aligned}$$

The observed value $W^+ = 42$ is much larger than the mean. We now expect that the data are statistically significant.

The *P*-value for our one-sided alternative is $P(W^+ \geq 42)$, calculated using the distribution of W^+ when the null hypothesis is true. Software gives a *P*-value of $P = 0.0098$.

CONCLUDE: The data give strong evidence (P<0.01$P < 0.01$) that neural activity is higher during the response to the visual stimulus (R) than just before (SA). This neuron’s receptive field is activated by the visual pattern presented.

APPLY YOUR KNOWLEDGE

27.14 Sitting versus squatting. Squatting for defecation continues to be the traditional position in most of Asia and Africa. In contrast, sitting on toilet seats has proliferated in Western counties since the 19th century with the development of the sewage system. This difference, along with differences in diet, might explain the higher rate of hemorrhoids and constipation in Western countries. A study asked if body position affects defecation in humans. Researchers recruited 6 healthy adult volunteers and took X-rays of their intestinal system to measure the anorectal angle (in degrees). Note that, as with a water hose, an angle of 180 degrees should provide the least amount of flow disruption. For each subject, an X-ray was taken once in a sitting position and once in a squatting position. Here are the findings:[10]

Subject	Sitting	Squatting
1	109	122
2	73	127
3	117	141
4	77	103
5	98	121
6	125	142

a. Find the differences within pairs, arrange them in order, and rank the absolute values. What is the signed rank statistic W^+?

b. If the null hypothesis (no difference in anorectal angle) is true, what are the mean and standard deviation of W^+? Does comparing W^+ with this mean lead to a tentative conclusion?

Borderlands/Alamy

27.15 Floral scents and learning. We hear that listening to Mozart improves students' performance on tests. In the EESEE case study "Floral Scents and Learning," investigators asked whether pleasant odors have a similar effect. Twenty-one subjects worked a paper-and-pencil maze while wearing a mask. The mask was either unscented or carried a floral scent. The response variable is the subjects' average time on three trials per condition. Each subject worked the maze with both masks, in a random order. The randomization is important because subjects tend to improve their times as they work a maze repeatedly. Table 27.1 gives the subjects' average times with each mask. Does the scent improve performance (that is, shorten the time needed to complete the maze)? A matched pairs *t* test works well and gives P=0.3652$P = 0.3652$. Let's compare the Wilcoxon signed rank test.

a. What are the ranks for the absolute values of the differences in Table 27.1? What is the value of W^+?

b. What would be the mean and standard deviation of W^+ if the null hypothesis (scent makes no difference) were true? Compare W^+ with this mean (in standard deviation units) to reach a tentative conclusion about significance.

TABLE 27.1 Average time (seconds) to complete a maze

Subject	Unscented	Scented	Difference	Subject	Unscented	Scentec
1	30.60	37.97	−7.37	12	58.93	83.50
2	48.43	51.57	−3.14	13	54.47	38.30
3	60.77	56.67	4.10	14	43.53	51.37

4	36.07	40.47	−4.40	15	37.93	29.33
5	68.47	49.00	19.47	16	43.50	54.27
6	32.43	43.23	−10.80	17	87.70	62.73
7	43.70	44.57	−0.87	18	53.53	58.00
8	37.10	28.40	8.70	19	64.30	52.40
9	31.17	28.23	2.94	20	47.37	53.63
10	51.23	68.47	−17.24	21	53.67	47.00
11	65.40	51.10	14.30			

The Normal approximation for W+W^+ The distribution of the signed rank statistic when the null hypothesis (no difference) is true becomes approximately Normal as the sample size becomes large. We can then use Normal probability calculations (with the continuity correction) to obtain approximate *P*-values for W^+. Let's see how this works in the receptive field example, even though n=9$n = 9$ is not a large sample.

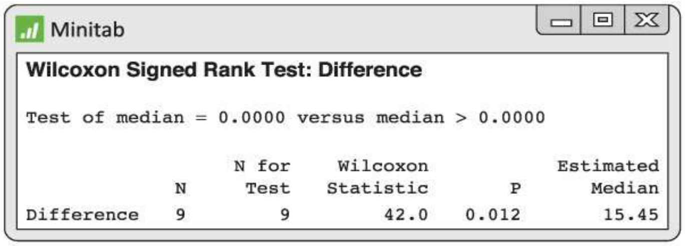
EXAMPLE 27.8 Visual receptive field: Normal approximation

For n=9$n = 9$ observations, we saw in Example 27.7 that μW+=22.5 $\mu_{W^+} = 22.5$ and that σW+=8.441$\sigma_{W^+} = 8.441$. We observed W+=42$W^+ = 42$, so the one-sided *P*-value is P(W+≥42)$P(W^+ \geq 42)$. The continuity correction calculates this as P(W+≥41.5)$P(W^+ \geq 41.5)$, treating the value W+=42$W^+ = 42$ as occupying the interval from 41.5 to 42.5. We find the Normal approximation for the *P*-value either from software or by standardizing and using the standard Normal table:

P(W+≥41.5)=PW+-22.58.441≥41.5-22.58.441=P(Z≥2.25)=0.0122

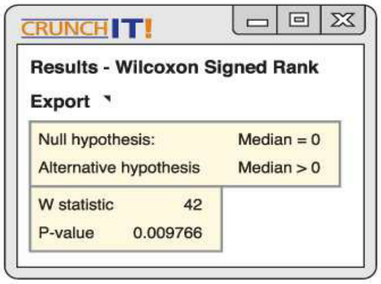
$$\begin{aligned} P(W^+ \geq 41.5) &= P\left(\frac{W^+ - 22.5}{8.441} \geq \frac{41.5 - 22.5}{8.441}\right) \\ &= P(Z \geq 2.25) \\ &= 0.0122 \end{aligned}$$

Figure 27.5 displays the output of four statistical programs. Minitab uses the Normal approximation and agrees with our calculation P=0.0122 $P = 0.0122$. SPSS also uses the Normal approximation but does not perform the continuity correction and gets a slightly different two-sided *P*-value. CrunchIt! uses the exact distribution of W^+ and finds that the exact one-sided *P*-value for the Wilcoxon signed rank test is P=0.0098$P = 0.0098$, as we reported in Example 27.7. The Normal approximation is quite close to this. R gives us the option to compute the *P*-value with either method. A matched pairs *t* test (not shown) finds P=0.007$P = 0.007$, but its validity depends on the distribution of the difference in neural activity. All tests tell us that there is strong evidence that this neuron’s receptive field is activated by the visual pattern presented.



Wilcoxon Signed Rank Test: Difference

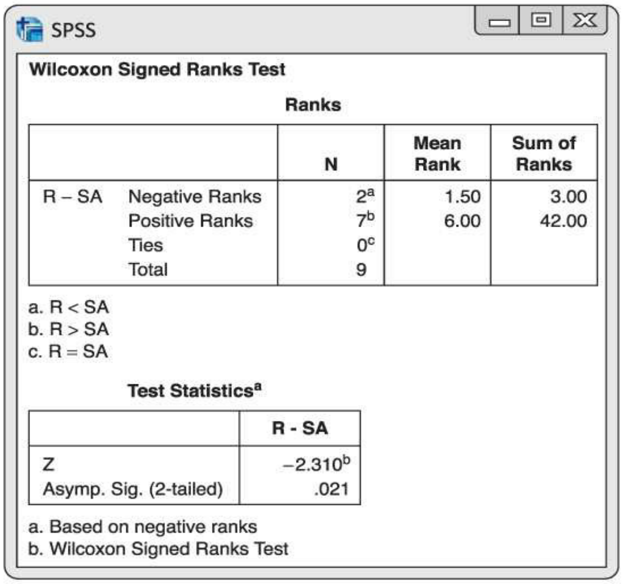
```
Test of median = 0.0000 versus median > 0.0000

                  N for   Wilcoxon            Estimated
             N    Test   Statistic      P       Median
Difference   9       9        42.0  0.012        15.45
```



Results - Wilcoxon Signed Rank

Null hypothesis:	Median = 0
Alternative hypothesis	Median > 0

W statistic	42
P-value	0.009766



Wilcoxon Signed Ranks Test

Ranks

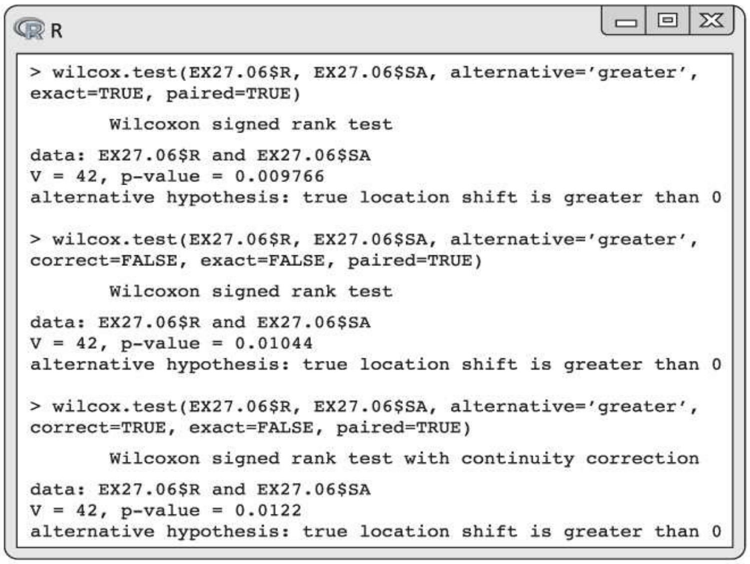
		N	Mean Rank	Sum of Ranks
R – SA	Negative Ranks	2[a]	1.50	3.00
	Positive Ranks	7[b]	6.00	42.00
	Ties	0[c]		
	Total	9		

a. R < SA
b. R > SA
c. R = SA

Test Statistics[a]

	R - SA
Z	−2.310[b]
Asymp. Sig. (2-tailed)	.021

a. Based on negative ranks
b. Wilcoxon Signed Ranks Test



```
> wilcox.test(EX27.06$R, EX27.06$SA, alternative='greater',
exact=TRUE, paired=TRUE)

       Wilcoxon signed rank test

data: EX27.06$R and EX27.06$SA
V = 42, p-value = 0.009766
alternative hypothesis: true location shift is greater than 0

> wilcox.test(EX27.06$R, EX27.06$SA, alternative='greater',
correct=FALSE, exact=FALSE, paired=TRUE)

       Wilcoxon signed rank test

data: EX27.06$R and EX27.06$SA
V = 42, p-value = 0.01044
alternative hypothesis: true location shift is greater than 0

> wilcox.test(EX27.06$R, EX27.06$SA, alternative='greater',
correct=TRUE, exact=FALSE, paired=TRUE)

       Wilcoxon signed rank test with continuity correction

data: EX27.06$R and EX27.06$SA
V = 42, p-value = 0.0122
alternative hypothesis: true location shift is greater than 0
```

Figure 27.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.5 Output from Minitab, CrunchIt!, SPSS, and R for the neural activity data of Example 27.6. CrunchIt! uses the exact method for calculating *P*, whereas Minitab and SPSS use a Normal approximation (one with and the other without a continuity correction). R gives the option of selecting the computation method (all three are included for comparison).

APPLY YOUR KNOWLEDGE

27.16 Sitting versus squatting: Normal approximation. Continue your work from Exercise 27.14. Use the Normal approximation with continuity correction to find the *P*-value for the signed rank test against a two-sided alternative. What do you conclude?

27.17 Sitting versus squatting: W+W^+ **versus** ***t*****.** Find the two-sided *P*-value for the matched pairs *t* test applied to the data in Exercise 27.14. The smaller *P*-value of *t* relative to W^+ means that *t* gives stronger evidence of the effect of sitting position. The *t* test takes advantage of assuming that the data are Normal, a considerable advantage for these very small samples. However, if you are not sure whether the data really are Normal, the nonparametric test is your only option.

27.18 Floral scents and learning: Normal approximation. Use the Normal approximation with continuity correction to find the *P*-value for the test in Exercise 27.15. Does the Wilcoxon signed rank test lead to essentially the same result as the P=0.3652$P = 0.3652$ for the *t* test?

27.19 Ancient air. Amber can preserve little samples of the past. Here are the percents of nitrogen found in the gas bubbles of 9 specimens of amber from the late Cretaceous era (75 to 95 million years ago):[11]

63.4 65.0 64.4 63.3 54.8 64.5 60.8 49.1 51.0

We wonder if ancient air differs significantly from the present atmosphere, which is 78.1% nitrogen.

a. Graph the data and comment on skewness and outliers. A rank test is appropriate.

b. We would like to test hypotheses about the median percent of nitrogen in ancient air (the population):

H0: median=78.1Ha: median≠78.1

$$H_0 : \text{median} = 78.1$$
$$H_a : \text{median} \neq 78.1$$

To do this, apply the Wilcoxon signed rank statistic to the differences between the observations and 78.1. (This is the one-sample version of the test.) What do you conclude?

Dealing with ties in the signed rank test Ties among the absolute differences are handled by assigning average ranks. A tie *within* a pair creates a difference of zero. Because these are neither positive nor negative, we drop such pairs from our sample. Ties within pairs simply reduce the number of observations, but ties among the absolute differences complicate finding a *P*-value. There is no longer a usable exact distribution for the signed rank statistic W^+, and the standard deviation σW+σ_{W^+} must be adjusted for the ties before we can use the Normal approximation. Software will do this. Here is an example.

EXAMPLE 27.9 Treatment for amyloidosis

STATE: Chronic rheumatoid arthritis can lead to secondary amyloidosis, a disabling and potentially fatal disease due to the buildup of amyloid proteins in specific organs. A Spanish study examined the efficacy and safety of anti–tumor necrosis factor in treating secondary amyloidosis affecting the kidneys.

Creatinine is a normal, nonprotein waste product of muscular function and is filtered out by the kidneys. The serum creatinine level is therefore a good marker of kidney function. In healthy individuals it remains fairly constant. Here are the serum creatinine levels (in milligrams per deciliters, mg/dl) before and after treatment of 14 female patients:[12]

Before	2.7	2.4	4.0	2.9	0.9	1.4	1.2	1.5	0.8	0.8	1.1	0.7	2.0
After	2.2	2.0	4.2	2.3	0.9	1.7	1.4	1.2	1.0	1.0	1.2	0.9	1.5
Diff.	−0.5	−0.4	0.2	−0.6	0.0	0.3	0.2	−0.3	0.2	0.2	0.1	0.2	−0.5

Negative differences represent a decrease in serum creatinine level.

Based on this sample, can we conclude that female patients given the anti-tumor necrosis factor will experience a change in their serum creatinine level?

PLAN: We would like to test the hypotheses that in women given the anti–tumor necrosis factor,

H0: serum creatinine levels are the same before and after treatmentHa: serum creatinine levels are systematically different after treatment

H_0 : serum creatinine levels are the same before and after treatment
H_a : serum creatinine levels are systematically different after treatment

SOLVE: A dotplot of the differences (Figure 27.6) shows a left-skew with one very low outlier. We will use the Wilcoxon signed rank test.

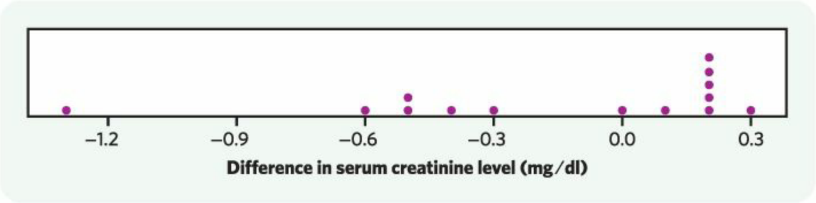
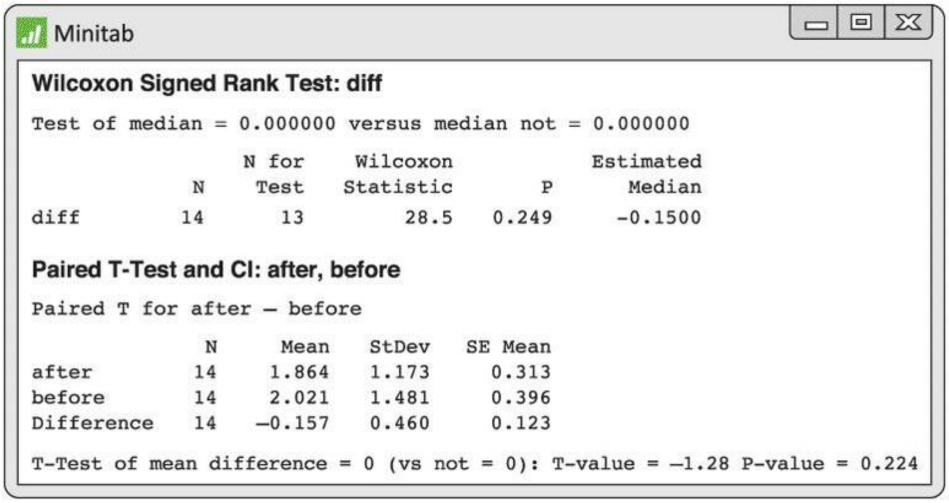


Figure 27.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.6 Dotplot of the differences in serum creatinine level before and after treatment, for Example 27.9.

Figure 27 displays the Minitab output for the serum creatinine data. The first query is the Wilcoxon test with its statistic W+=28.5$W^+ = 28.5$ and a two-sided *P*-value P=0.249$P = 0.249$. The second query is the output of a matched pairs *t* test, for which P=0.224$P = 0.224$. The two *P*-values are somewhat similar and the practical conclusion is the same. The nonparametric test is more reliable for samples with outliers.



Wilcoxon Signed Rank Test: diff

```
Test of median = 0.000000 versus median not = 0.000000

                  N for    Wilcoxon              Estimated
              N    Test   Statistic        P        Median
diff         14      13        28.5    0.249       -0.1500
```

Paired T-Test and CI: after, before

```
Paired T for after – before

               N     Mean    StDev   SE Mean
after         14    1.864    1.173     0.313
before        14    2.021    1.481     0.396
Difference    14   –0.157    0.460     0.123

T-Test of mean difference = 0 (vs not = 0): T-value = –1.28 P-value = 0.224
```

Figure 27.7

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.7 Output from Minitab for the serum creatinine data of Example 27.9. Because there are ties, a Normal approximation must be used for the Wilcoxon signed rank test. The results of a paired *t* test are provided for comparison.

CONCLUDE: These data give no evidence for a systematic change in serum creatinine level following treatment. Kidney function appears stable.

Let’s see where the value W+=28.5$W^+ = 28.5$ came from. One difference was equal to zero and thus was omitted. The absolute values of the remaining differences, with boldface indicating those that were negative, are

0.5 **0.4** 0.2 **0.6** 0.3 0.2 **0.3** 0.2 0.2 0.1 0.2 **0.5** **1.3**

Arrange these in increasing order and assign ranks, keeping track of which values were originally negative. Tied values receive the average of their ranks.

Abs.value	0.1	0.2	0.2	0.2	0.2	0.2	**0.3**	0.3	**0.4**	**0.5**	**0.5**	**0.6**	**1.3**
Rank	1	4	4	4	4	4	**7.5**	7.5	**9**	**10.5**	**10.5**	**12**	**13**

The Wilcoxon signed rank statistic is the sum W+=28.5$W^+ = 28.5$ of the ranks of the positive differences. (We could equally well use the sum of the ranks of the negative differences.)

APPLY YOUR KNOWLEDGE

27.20 Does nature heal best? Table 25.2 (page 644) gives data on the healing rate (micrometers per hour) of the skin of newts under two conditions. This is a matched pairs design, with the body's natural electric field for one limb (control) and half the natural value for another limb of the same newt (experimental). We want to know if the healing rates are systematically different under the two conditions. You decide to use a rank test.

a. There are several ties among the absolute differences. Find the ranks and give the value of the signed rank statistic W^+.

b. Use software to find the *P*-value. Give a conclusion. Be sure to include a description of what the data show in addition to the test results.

27.21 Sweetening colas. Cola makers test new recipes for loss of sweetness during storage. Trained tasters rate the sweetness before and after storage. Here are the sweetness losses (sweetness before storage minus sweetness after storage) found by 10 tasters for one new cola recipe:

2.0 0.4 0.7 2.0 −0.4 2.2 −1.3 1.2 1.1 2.3

Are these data good evidence that the cola lost sweetness?

a. These data are the differences from a matched pairs design. State hypotheses in terms of the median difference in the population of all tasters, carry out a test, and give your conclusion.

b. Software gives P=0.0123$P = 0.0123$ for the one-sample *t* test for these data. How does this compare with your result from a? What are the hypotheses for the *t* test? What conditions must be met for the *t* test? What conditions must be met for the Wilcoxon test?

COMPARING SEVERAL SAMPLES: THE KRUSKAL-WALLIS TEST

We have now considered alternatives to the paired-sample and two-sample *t* tests for comparing the magnitude of responses to two treatments. To compare mean responses for more than two treatments, we use one-way analysis of variance (ANOVA) if the distributions of the responses to each treatment are at least roughly Normal and have similar spreads. What can we do when these distribution requirements are violated?

EXAMPLE 27.10 Weeds among the corn

STATE: Lamb's-quarter is a common weed that interferes with the growth of corn. A researcher planted corn at the same rate in 16 small plots of ground, then randomly assigned the plots to four groups. He weeded the plots by hand to allow a fixed number of lamb's-quarter plants to grow in each meter of corn row. These numbers were 0, 1, 3, and 9 in the four groups of plots. No other weeds were allowed to grow, and all plots received identical treatment except for the weeds. Here are the yields of corn (bushels per acre) in each of the plots:[13]

Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield	Weeds per meter	Corn yield
0	166.7	1	166.2	3	158.6	9	162.8
0	172.2	1	157.3	3	176.4	9	142.4
0	165.0	1	166.7	3	153.1	9	162.7
0	176.9	1	161.1	3	156.0	9	162.4

Do yields change as the number of weeds changes?

PLAN: Do data analysis to see how the yields change. Test the null hypothesis "no difference in the distribution of yields" against the alternative that the groups do differ.

SOLVE (first steps): The summary statistics are

Weeds	*n*	Median	Mean	Std. dev.

0	4	169.45	170.200	5.422
1	4	163.65	162.825	4.469
3	4	157.30	161.025	10.493
9	4	162.55	157.575	10.118

The mean yields do go down as more weeds are added. ANOVA tests whether the differences are statistically significant. Can we safely use ANOVA? Outliers are presentin the yields for 3 and 9 weeds per meter. The outliers explain the differences between the means and the medians. They are the correct yields for their plots, so we cannot remove them. Moreover, the sample standard deviations do not quite satisfy our rule of thumb for ANOVA that the largest should not exceed twice the smallest. We may prefer to use a nonparametric test.

Hypotheses and conditions for the Kruskal-Wallis test The ANOVA *F* test concerns the means of the several populations represented by our samples. For Example 27.10 the ANOVA hypotheses are

H0: μ0=μ1=μ3=μ9Ha: not all four means are equal

$$H_0 : \mu_0 = \mu_1 = \mu_3 = \mu_9$$
$$H_a : \text{not all four means are equal}$$

For example, μ0μ_0 is the mean yield in the population of all corn planted under the conditions of the experiment with no weeds present. The data should consist of four independent random samples from the four populations, all Normally distributed with the same standard deviation.

The *Kruskal-Wallis test* is a rank test that can replace the ANOVA *F* test. The condition about data production (independent random samples from each population) remains important, but we can relax the Normality condition. We assume only that the response has a continuous distribution in each population. The hypotheses tested in our example are

H0:yields have the same distribution in all groupsHa:yields are systematically higher in some groups than in others

$$H_0 : \text{yields have the same distribution in all groups}$$
$$H_a : \text{yields are systematically higher in some groups than in others}$$

If all the population distributions have the same shape (Normal or not), these hypotheses take a simpler form. The null hypothesis is that all four populations have the same *median* yield. The alternative hypothesis is that not all four median yields are equal. The different standard deviations suggest that the four distributions in Example 27.10 do *not* all have the same shape.

The Kruskal-Wallis test statistic Recall the analysis of variance idea: We write the total observed variation in the responses as the sum of two parts, one measuring variation among the groups (sum of squares for groups, SSG) and one measuring variation among individual observations within the same group (sum of squares for error, SSE). The ANOVA *F* test rejects the null hypothesis that the mean responses are equal in all groups if SSG is large relative to SSE.

The idea of the Kruskal-Wallis rank test is to rank all the responses from all groups together and then apply one-way ANOVA to the ranks rather than to the original observations. If there are *N* observations in all, the ranks are always the whole numbers from 1 to *N*. The total sum of squares for the ranks is therefore a fixed number no matter what the data are. So we do not need to look at both SSG and SSE. Although it isn't obvious without some unpleasant algebra, the Kruskal-Wallis test statistic is essentially just SSG for the ranks. We give the formula, but you should rely on software to do the arithmetic. When SSG is large, that is evidence that the groups differ.

THE KRUSKAL-WALLIS TEST

Draw independent SRSs of sizes n1,n2,…,nk$n_1, n_2, \ldots, n_k$ from *k* populations. There are *N* observations in all. Rank all *N* observations, and let RiR_i be the sum of the ranks for the *i*th sample. The **Kruskal-Wallis statistic** is

H=12N(N+1)∑Ri2ni-3(N+1)

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

When the sample sizes nin_i are large and all *k* populations have the same continuous distribution, *H* has approximately the chi-square distribution with k-1$k-1$ degrees of freedom.

The **Kruskal-Wallis test** rejects the null hypothesis that all populations have the same distribution when *H* is large.

We now see that, like the Wilcoxon rank sum statistic, the Kruskal-Wallis statistic is based on the sums of the ranks for the groups we are comparing. The more different these sums are, the stronger is the evidence that responses are systematically larger in some groups than in others.

The exact distribution of the Kruskal-Wallis statistic *H* under the null hypothesis depends on all the sample sizes n1n_1 to nk,$n_k,$ so tables are awkward. The calculation of the exact distribution is so time-consuming for all but the smallest problems that even most statistical software uses the chi-square approximation to obtain *P*-values. As usual, there is no usable exact distribution when there are ties among the responses. We again assign average ranks to tied observations.

EXAMPLE 27.11 Weeds among the corn, continued

SOLVE (inference): In Example 27.10 there are k=4$k = 4$ populations and N=16$N = 16$ observations. The sample sizes are equal, ni=4$n_i = 4$. The 16 observations arranged in increasing order, with their ranks, are

Yield	142.4	153.1	156.0	157.3	158.6	161.1	162.4	162.7
Rank	1	2	3	4	5	6	7	8

Yield	162.8	165.0	166.2	166.7	166.7	172.2	176.4	176.9
Rank	9	10	11	12.5	12.5	14	15	16

There is one pair of tied observations. The ranks for each of the four treatments are

Weeds	Ranks				Sum of ranks
0	10	12.5	14	16	52.5
1	4	6	11	12.5	33.5
3	2	3	5	15	25.0
9	1	7	8	9	25.0

The Kruskal-Wallis statistic is therefore

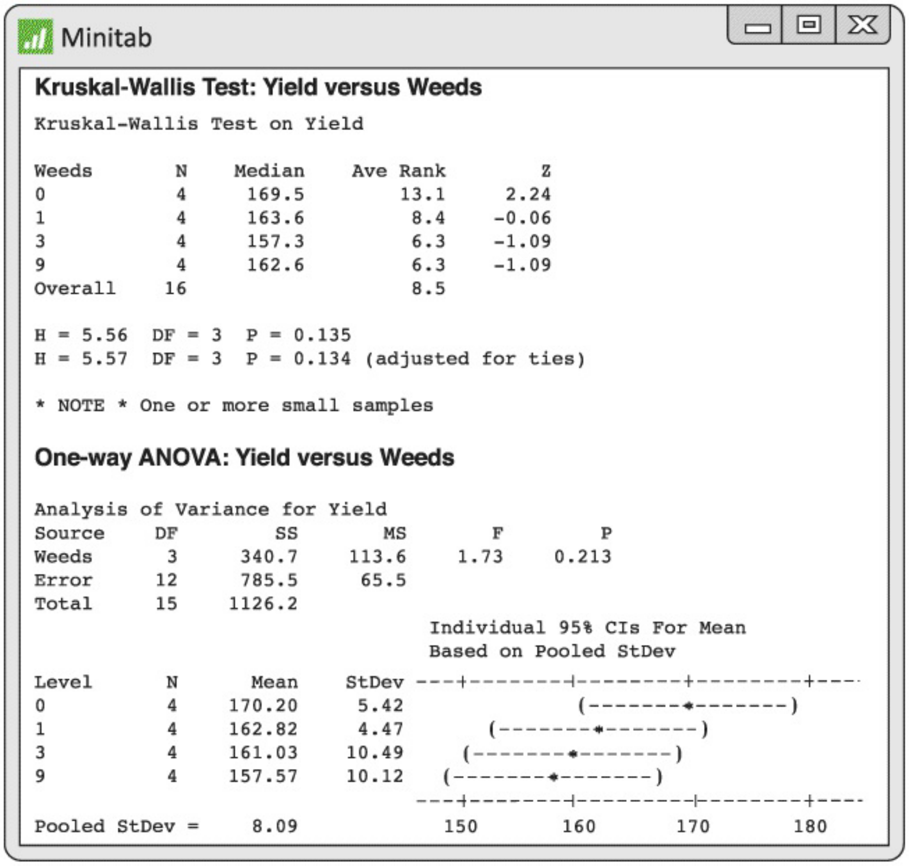
H=12N(N+1)∑Ri2ni-3(N+1)=12(16)(17)52.524+33.524+2524+2524-(3)(17)=12272(1282.125)-51=5.56

$$
\begin{aligned}
H &= \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1) \\
&= \frac{12}{(16)(17)} \left(\frac{52.5^2}{4} + \frac{33.5^2}{4} + \frac{25^2}{4} + \frac{25^2}{4} \right) - (3)(17) \\
&= \frac{12}{272}(1282.125) - 51 \\
&= 5.56
\end{aligned}
$$

Referring to the table of chi-square critical points (Table D) with df =$=$ 3, we see that the *P*-value lies in the interval 0.10<P<0.15$0.10 < P < 0.15$.

CONCLUDE: This small experiment does not provide convincing evidence that weeds decrease corn yield.

Figure 27.8 displays the Minitab output for both the Kruskal-Wallis test and the ANOVA test. Minitab finds that H=5.56$H = 5.56$ and gives P=0.135$P = 0.135$. Minitab also gives the results of an adjustment that makes the chi-square approximation more accurate when there are ties. For these data, the adjustment has no practical effect. It would be important if there were many ties. A very lengthy computer calculation shows that the exact *P*-value is P=0.1299$P = 0.1299$. The chi-square approximation is quite accurate.



Kruskal-Wallis Test: Yield versus Weeds

```
Kruskal-Wallis Test on Yield

Weeds       N    Median    Ave Rank       Z
0           4     169.5        13.1    2.24
1           4     163.6         8.4   -0.06
3           4     157.3         6.3   -1.09
9           4     162.6         6.3   -1.09
Overall    16                   8.5

H = 5.56  DF = 3  P = 0.135
H = 5.57  DF = 3  P = 0.134 (adjusted for ties)

* NOTE * One or more small samples
```

One-way ANOVA: Yield versus Weeds

```
Analysis of Variance for Yield
Source    DF         SS       MS       F        P
Weeds      3      340.7    113.6    1.73    0.213
Error     12      785.5     65.5
Total     15     1126.2
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level      N      Mean     StDev ---+---------+---------+---------+---
0          4     170.20     5.42              (--------*--------)
1          4     162.82     4.47       (--------*--------)
3          4     161.03    10.49     (--------*--------)
9          4     157.57    10.12   (--------*--------)
                                 ---+---------+---------+---------+----
Pooled StDev =    8.09             150       160       170       180
```

Figure 27.8

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 27.8 Minitab output for the corn yield data of Example 27.10. For comparison, both the Kruskal-Wallis test and the one-way ANOVA are shown.

The ANOVA *F* test gives F=1.73$F = 1.73$ with P=0.213$P = 0.213$. Although the practical conclusion is the same, ANOVA and Kruskal-Wallis do not agree closely in this example. The rank test is more reliable for these small samples with outliers.

APPLY YOUR KNOWLEDGE

27.22 Which color attracts beetles best? To detect the presence of harmful insects in farm fields, we can put up boards covered with a sticky material and examine the insects trapped on the boards. Experimenters placed six boards of each of four colors at random locations in a field of oats and measured the number of cereal leaf beetles trapped. Here are the data:[14]

Color	Beetles trapped					
Blue	16	11	20	21	14	7
Green	37	32	20	29	37	32
White	21	12	14	17	13	20
Yellow	45	59	48	46	38	47

The samples are small. If you have no reasons to believe that the data are Normally distributed (say, from similar studies with large sample sizes), it would be safer to apply a nonparametric test.

a. Find the median number of beetles trapped by boards of each color. Which colors appear more effective?

b. What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?

c. What are *k*, the nin_i, and *N*? Arrange the counts in order and assign ranks. Be careful about ties.

d. Calculate the Kruskal-Wallis statistic *H*. How many degrees of freedom should you use for the chi-square approximation to its null distribution? Use the chi-square table to give an approximate *P*-value. What does the test lead you to conclude?

27.23 Logging in the rain forest: species richness. Table 24.2 (page 616) contains data comparing the number of trees and number of tree species in plots of land in a tropical rain forest that had never been logged with similar plots nearby that had been logged 1 year earlier and 8 years earlier. The third response variable is species richness, the number of tree species divided by the number of trees. There are low outliers in the data, and a histogram of the ANOVA residuals shows outliers as well. Because of lack of Normality and small samples, we may prefer the Kruskal-Wallis test.

a. Make a graph to compare the distributions of richness for the three groups of plots. Also give the median richness for the three groups.

b. Use the Kruskal-Wallis test to compare the distributions of richness. State hypotheses, the test statistic and its *P*-value, and your conclusions.

27.24 More rain for California? Exercise 24.29 (page 629) describes an experiment that examines the effect on plant biomass in plots of California grassland randomly assigned to receive added water in the winter, added water in the spring, or no added water. The experiment continued for several years. Here are data for 2004 (mass in grams per square meter):

Winter	Spring	Control
254.6453	517.6650	178.9988
233.8155	342.2825	205.5165
253.4506	270.5785	242.6795
228.5882	212.5324	231.7639
158.6675	213.9879	134.9847
212.3232	240.1927	212.4862

The sample sizes are small and the data contain some possible outliers. We will apply a nonparametric test.

a. Examine the data. Show that the conditions for ANOVA (page 618) are not met. What appear to be the effects of extra water in winter or spring?

b. What hypotheses does ANOVA test? What hypotheses does Kruskal-Wallis test?

c. What are *k*, the nin_i, and *N*? Arrange the data in order and assign ranks.

d. Calculate the Kruskal-Wallis statistic *H*. How many degrees of freedom should you use for the chi-square approximation to its null distribution? Use the chi-square table to give an approximate *P*-value. What does the test lead you to conclude?

CHAPTER 27 SUMMARY

- **Nonparametric tests** do not require any specific form for the distributions of the populations from which our samples come.
- **Rank tests** are nonparametric tests based on the **ranks** of observations, their positions in the list ordered from smallest (rank 1) to largest. Tied observations receive the average of their ranks. Use rank tests when the data come from random samples or randomized comparative experiments and the populations have continuous distributions.
- The **Wilcoxon rank sum test** compares two distributions to assess whether one has systematically larger values than the other. The Wilcoxon test is based on the **Wilcoxon rank sum statistic** W**,** which is the sum of the ranks of one of the samples. The Wilcoxon test can replace the **two-sample** t **test.** Software may perform the **Mann-Whitney test,** another form of the Wilcoxon test.
- ***P*-values** for the rank sum test are based on the sampling distribution of the rank sum statistic W when the null hypothesis (no difference in distributions) is true. You can find P-values from special tables, software, or a Normal approximation (with continuity correction).
- The **Wilcoxon signed rank test** applies to matched pairs studies. It tests the null hypothesis that there is no systematic difference within pairs against alternatives that assert a systematic difference (either one-sided or two-sided).
- The test is based on the **Wilcoxon signed rank statistic** W^+**,** which is the sum of the ranks of the positive (or negative) differences when we rank the absolute values of the differences. The **matched pairs** t **test** is an alternative test in this setting.
- ***P*-values** for the signed rank test are based on the sampling distribution of W^+ when the null hypothesis is true. You can find P-values from special tables, software, or a Normal approximation (with continuity correction).

- The **Kruskal-Wallis test** compares several populations on the basis of independent random samples from each population. This is the **one-way analysis of variance** setting.
- The null hypothesis for the Kruskal-Wallis test is that the distribution of the response variable is the same in all the populations. The alternative hypothesis is that responses are systematically larger in some populations than in others.
- The **Kruskal-Wallis statistic** *H* can be viewed in two ways. It is essentially the result of applying one-way ANOVA to the ranks of the observations. It is also a comparison of the sums of the ranks for the several samples.
- When the sample sizes are not too small and the null hypothesis is true, the Kruskal-Wallis test statistic for comparing *k* populations has approximately the chi-square distribution with k-1$k - 1$ degrees of freedom. We use this approximate distribution to obtain *P*-values.

STATISTICS IN SUMMARY

Here are the most important skills you should have acquired from reading this chapter.

A. RANKS

1. Assign ranks to a moderate number of observations. Use average ranks if there are ties among the observations.
2. From the ranks, calculate the rank sums when the observations come from two or several samples.

B. RANK TEST STATISTICS

1. Determine which of the rank sum tests is appropriate in a specific problem setting.
2. Calculate the Wilcoxon rank sum statistic W from ranks for two samples, the Wilcoxon signed rank sum statistic W^+ for matched pairs, and the Kruskal-Wallis statistic H for two or more samples.
3. State the hypotheses tested by each of these statistics in specific problem settings.
4. Determine when it is appropriate to state the hypotheses for W and H in terms of population medians.

C. RANK TESTS

1. Use software to carry out any of the rank tests. Combine the test with data description and give a clear statement of findings in specific problem settings.
2. Use the Normal approximation with continuity correction to find approximate P-values for W and W^+. Use a table of chi-square critical values to approximate the P-value of H.

THIS CHAPTER IN CONTEXT

In Chapter 17 we described the one-sample *t* procedures used to perform statistical inference for one population mean *μ*. This was followed in Chapter 18 by the two-sample *t* procedures used to perform statistical inference for the difference μ1-μ2$\mu_1 - \mu_2$ between the means of two distinct populations. In Chapter 24 we described the one-way analysis of variance *F* test, a method used to compare the means μ1,μ2,…,μk$\mu_1, \mu_2, \ldots, \mu_k$ of several populations while avoiding the problems associated with multiple comparisons.

The *t* procedures and ANOVA are used when the response variable is quantitative, and they assume that the underlying populations are Normally distributed. So, checking the conditions for inference typically requires that we examine the data for evidence of skew or outliers, as seen in Chapters 1 and 2. The *t* procedures and ANOVA, fortunately, are robust with respect to this assumption of Normality when the sample sizes are large enough.

In this chapter we examine nonparametric inference procedures equivalent to the one-sample and matched pairs *t* tests, the two-sample *t* test, and the ANOVA *F* test. Respectively, these are the Wilcoxon signed rank test, the Wilcoxon rank sum test (also called the Mann-Whitney test), and the Kruskal-Wallis test. These procedures rely on the ranking of the data rather than the actual numerical values. Normality of the response variable is not required. Therefore, these nonparametric procedures are great alternatives when we cannot be sure that the response variable is truly continuous or Normally distributed, and the samples sizes are too small to rely on the robustness of the *t* and *F* methods.

CHECK YOUR SKILLS

27.25 A study of blood pressure and age compares the blood pressures of men in three age groups: less than 30 years, 30 to 55 years, and over 55 years. To analyze the data you would use the

a. Wilcoxon rank sum test.
b. Wilcoxon signed rank test.
c. Kruskal-Wallis test.

27.26 The previous study also compares the blood pressures of men to those of women for the 30 to 55 age group. To analyze the data you would use the

a. Wilcoxon rank sum test.
b. Wilcoxon signed rank test.
c. Kruskal-Wallis test.

27.27 You interview 75 students in their freshman year and again in their senior year. Each interview includes information about height and weight, allowing the computation of each student's body mass index (BMI). To assess whether there has been a significant change from freshman to senior year in their BMIs, you would use the

a. Wilcoxon rank sum test.
b. Wilcoxon signed rank test.
c. Kruskal-Wallis test.

27.28 When some plants are attacked by leaf-eating insects, they release chemical compounds that repel the insects. Here are data on emissions of one compound by 6 plants attacked by leaf bugs and by 6 plants in an undamaged control group:

Control group	14.4	15.2	12.6	11.9	5.1	8.0
Attacked group	10.6	15.3	25.2	19.8	17.1	14.6

The rank sum *W* for the control group is

a. 21.
b. 26.
c. 52.

27.29 If there is no difference in emissions between the attacked group and the control group, the mean of W in the previous exercise is

a. 39.
b. 78.
c. 6.2.

27.30 Suppose that the 12 observations in Exercise 27.28 were

Control group	14.4	15.2	12.6	11.9	5.1	8.0
Attacked group	12.6	15.3	25.2	19.8	17.1	14.4

The rank sum for the control group would then be

a. 21.
b. 25.
c. 26.

27.31 Interview 10 young married couples, wife and husband separately. One question asks how important the attractiveness of their spouse is to them on a scale of 1 to 10. Here are the responses:

	Couple									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Husband	7	7	7	3	9	5	10	6	6	7
Wife	4	2	5	2	2	2	4	7	1	5

The Wilcoxon signed rank statistic W^+ (based on husband's score minus wife's score) is

a. 51.
b. 53.5.
c. 54.

27.32 If husbands and wives don't differ in how important the attractiveness of their spouse is, the mean of W^+ in the previous exercise is

a. 27.5.
b. 55.
c. 105.

27.33 Suppose that the responses in Exercise 27.31 were

	Couple									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Husband	7	7	7	3	9	5	10	6	6	5
Wife	4	2	5	3	2	2	4	7	1	5

The Wilcoxon signed rank statistic W^+ (based on husband's score minus wife's score) is now

a. 35.
b. 36.
c. 52.

27.34 You compare the LDL ("bad") cholesterol levels of 4 college freshmen, 5 sophomores, 6 juniors, and 7 seniors. If the four LDL cholesterol distributions are the same, the Kruskal-Wallis statistic H has approximately a chi-square distribution. The degrees of freedom are

a. 3.
b. 4.
c. 18.

CHAPTER 27 EXERCISES

One of the rank tests discussed in this chapter is appropriate for each of the following exercises. Follow the **Plan, Solve,** *and* **Conclude** *parts of the four-step process in your answers. It may be helpful to restate in your own words the* **State** *information given in the exercise.*

27.35 Effectiveness of subliminal messages. A “subliminal” message is below our threshold of awareness but may nonetheless influence us. Can subliminal messages help students learn math? Table 27.2 gives the pretest and posttest scores for two groups of students taking a program to improve their basic mathematics skills.[15] In addition to the math program, all students received a daily subliminal message, flashed on a screen too rapidly to be consciously read. The treatment group of 10 students (chosen at random) was exposed to “Each day I am getting better in math.” The control group of 8 students was exposed to a neutral message, “People are walking on the street.” Is there evidence of significant improvement within each group? Did the treatment group show significantly greater improvement than the control group?

TABLE 27.2 **Mathematics skills scores before and after a subliminal message**

Treatment group		Control group	
Before	**After**	**Before**	**After**
18	24	18	29
18	25	24	29
21	33	20	24
18	29	18	26
18	33	24	38
20	36	22	27
23	34	15	22
23	36	19	31
21	34		
17	27		

27.36 Diet and reproduction in fruit flies. Exercise 27.5 introduced an experiment about fruit fly reproduction. The researchers also studied how diet composition affects their reproductive output. Wild-type fruit flies were assigned to receive varying amounts of yeast supplements (in milligrams, mg, per day), a protein-rich food, in their diet for two weeks, after which the number of eggs produced per female was assessed. Here are the results:

0 mg	7.55	6.10	9.25	5.40	3.75
1 mg	10.75	18.15	18.70	16.95	11.15
3 mg	18.35	39.25	44.60	42.80	24.20
7 mg	40.35	67.95	48.50	64.25	60.35

Does reproductive output differ based on diet richness?

27.37 Fighting cancer. Lymphocytes (white blood cells) play an important role in defending our bodies against tumors and infections. Can lymphocytes be genetically modified to recognize and destroy cancer cells? In one study of this idea, modified cells were infused into 11 patients with metastatic melanoma (serious skin cancer) that had not responded to existing treatments. Here are data for an "ELISA" test for the presence of cells that trigger an immune response, in counts per 100,000 cells before and after infusion.[16] High counts suggest that infusion had a beneficial effect.

Patient	1	2	3	4	5	6	7	8	9	10	11
Before	14	0	1	0	0	0	0	20	1	6	0
After	41	7	1	215	20	700	13	530	35	92	108

What do you conclude about the effect of infusing modified cells on the ELISA count?

27.38 Canine compulsive disorder. Obsessive compulsive disorder (OCD) is a human anxiety disorder with a poorly understood origin. Canine compulsive disorder (CCD) is an analogous behavioral disorder affecting dogs that may provide an animal model for OCD. Using magnetic resonance imaging, researchers compared the brain structures of

Doberman dogs with and without CCD. Here are standardized fractional anisotropy values in the corpus callosum, indicative of fiber density in this white-matter structure connecting the cerebral hemispheres:[17]

Control	0.47	0.46	0.38	0.27	0.26	0.27	0.24	0.19
CCD	0.73	0.62	0.61	0.49	0.45	0.37	0.3	0.32

Is there significant evidence that fractional anisotropy values in the corpus callosum are systematically larger in dogs with CCD or in those without CCD?

27.39 Nematodes and plant growth. A botanist prepares 16 identical planting pots and then introduces different numbers of nematodes (microscopic worms) into the pots. A tomato seedling is transplanted into each pot. Here are data on the increase in height of the seedlings (in centimeters) 16 days after planting:[18]

Nematodes	**Seedling growth**			
0	10.8	9.1	13.5	9.2
1,000	11.1	11.1	8.2	11.3
5,000	5.4	4.6	7.4	5.0
10,000	5.8	5.3	3.2	7.5

Do nematodes in soil affect plant growth?

27.40 Evolution in bacteria. Can bacteria evolve a preference for the pH of their environment? An evolutionary biologist took lines of *E. coli* bacteria grown and kept at neutral pH 7.2 and grew them for 2000 generations (about 300 days) at a stressful acidic pH of 5.5. The ancestor bacteria (kept frozen all that time) and the “acid-evolved” bacteria were then grown together at pH 5.5 to compute a relative fitness score. (A score of 1 indicates equal growth, a larger score indicates that the evolved bacteria grow better than their ancestors.) A control group consisting of some of the original bacteria was grown for 2000 generations in a neutral pH, and a relative fitness score (relative to the ancestor line) was later obtained at pH 5.5. Here are the scores for the acid-evolved and the control treatments from 6 replications of the experiment:[19]

Acid-evolved	1.24	1.22	1.23	1.24	1.18	1.09
Control	1.10	1.02	0.99	1.04	1.08	1.12

Do the data provide evidence that *E. coli* can evolve adaptations to an acidic pH over the course of 2000 generations? That is, is there evidence that the acid-evolved bacteria have higher relative fitness scores than control bacteria grown at a neutral pH?

27.41 A treatment for progeria. In Example 17.4 (page 431) we examined the effect of the drug lonafarnib on pulse wave velocity (PWV) in children diagnosed with progeria, a rare genetic condition that produces rapid aging in children. PWV is the standard measure of vascular stiffness, an important factor in cardiovascular health. Table 17.1 (page 431) gives the PWV values (in meters per second, m/s) at the beginning (“Untreated”) and at the end (“Treated”) of the two-year study for each child. Use a nonparametric method to determine whether PWV in children with progeria is significantly lower after treatment with lonafanib.

27.42 Vaccine protection against SIV. A research team investigated the protective effect of 4 vaccine variants against the simian immunodeficiency virus (SIV) in rhesus monkeys. V2-specific antibodies provide an indirect assessment of protection against SIV infection. Here are V2-specific antibody productions measured by surface plasmon resonance (in response units) for monkeys assigned to a vaccine variant or to a sham (fake) vaccine:[20]

Vaccine 1	Vaccine 2	Vaccine 3	Vaccine 4	Sham
100.1	83.5	171.6	196.1	11.4
53.9	73.0	118.9	154.3	4.8
25.3	32.3	111.1	119.4	6.3
17.9	14.5	103.4	120.2	4.0
22.1	15.3	93.3	77.9	0.1
14.0	1.3	88.6	65.5	1.7
4.3	0.6	77.0	74.0	1.3
10.9	0.6	60.0	68.6	1.3

a. Examine the 5 samples. Are the data appropriate for ANOVA?
b. Use the Kruskal-Wallis test to determine whether the data provide evidence that V2-specific antibody production is systematically larger with some treatments than with others.

27.43 Cicadas as fertilizer? Every 17 years, swarms of cicadas emerge from the ground in the eastern United States, live for about six weeks, then die. There are so many cicadas that their dead bodies can serve as fertilizer. In an experiment, a researcher added cicadas under some plants in a natural plot of bellflowers on the forest floor, leaving other plants undisturbed. Here are the seed masses (in milligrams) produced by 39 cicada-fertilized plants and by 33 undisturbed (control) plants:[21]

Cicada plants				**Control plants**			
0.237	0.277	0.241	0.142	0.212	0.188	0.263	0.253
0.109	0.209	0.238	0.277	0.261	0.265	0.135	0.170
0.261	0.227	0.171	0.235	0.203	0.241	0.257	0.155
0.276	0.234	0.255	0.296	0.215	0.285	0.198	0.266
0.239	0.266	0.296	0.217	0.178	0.244	0.190	0.212
0.238	0.210	0.295	0.193	0.290	0.253	0.249	0.253
0.218	0.263	0.305	0.257	0.268	0.190	0.196	0.220
0.351	0.245	0.226	0.276	0.246	0.145	0.247	0.140
0.317	0.310	0.223	0.229	0.241			
0.192	0.201	0.211					

Do the data show that dead cicadas increase seed mass?

27.44 More on visual receptive fields. Example 27.6 describes a neuron in the visual cortex stimulated by the presentation of a visual pattern on a screen. By using 3D glasses that introduce a visual disparity between the two retinas, that pattern can be made to appear as if it is floating in front of or behind the screen. Some neurons in the primary visual cortex specialize in the recognition of such 3D clues. Here is the response of the neuron from Example 27.6 to a visual pattern with varying amounts of disparity presented in random order:

Disparity	**Neural response (in spikes per second)**

−0.30	13.3	3.3	46.7	0.0	13.3	60.0	0.0	53.3	6.7
−0.15	13.3	6.7	6.7	0.0	6.7	6.7	10.0	13.3	0.0
0.00	16.7	20.0	23.3	16.7	56.7	0.0	26.7	36.7	10.0
0.15	86.7	3.3	46.7	86.7	50.0	56.7	80.0	73.3	106.7
0.30	53.3	23.3	0.0	13.3	0.0	23.3	0.0	0.0	13.3
0.60	23.3	40.0	20.0	0.0	23.3	6.7	6.7	23.3	6.7

Neurons specializing in 3D perception from retinal disparity respond differentially to patterns presented with different amounts of disparity. Is this such a neuron? What does it respond to best? (A positive disparity corresponds to the impression of a far object.)

27.45 Smoking during pregnancy. Cigarette labels warn pregnant women against smoking. Does nicotine actually reach the fetus, crossing the protective placental barrier? Researchers selected consecutive pregnant women delivering at an Egyptian hospital and categorized them as active smokers, passive smokers, or nonsmokers. They then analyzed the newborns' meconium for cotinine content, the metabolized form of nicotine. Meconium is a newborn's first stool right after birth and is a good biological marker for fetal exposure to drugs or other chemical agents. Here are the meconium cotinine levels (in nanograms per milliliter, ng/ml) for the 3 groups:[22]

Active smokers	490	418	405	328	700	292	295	272	240	232
Passive smokers	254	219	287	257	271	282	148	273	350	293
Nonsmokers	158	163	153	207	211	159	199	187	200	213

a. Examine the 3 samples. What are the overall shapes of the distributions? Are there outliers? What are the sample standard deviations? Explain why ANOVA cannot be used safely on these data.
b. Use a nonparametric method to determine whether meconium cotinine levels systematically change with the mother's smoking status.

27.46 Nanomedicine. Researchers examined a new treatment for advanced ovarian cancer in a mouse model. They created a nanoparticle-based delivery system for a suicide gene therapy to be delivered directly to the tumor cells. The grafted tumors were injected either with the new treatment or with only some buffer solution to serve as a comparison. The data below give the increase in size of the tumor after two weeks in 20 mice. A 1 represents no change, and a 2 represents a doubling in volume of the tumor.[23]

Buffer solution									
9.1	8.1	7.8	7.0	6.8	5.4	5.4	4.1	3.8	3.3
Nanoparticle-delivered gene therapy									
4.1	3.5	2.1	2.1	1.8	1.8	1.4	1.2	1.1	1.1

Do the data indicate that the nanoparticle-based delivery system is more effective than a placebo in slowing ovarian tumor growth in mice?

27.47 Ink toxicity. The National Toxicology Program evaluates the toxicity of chemicals found in manufacturing, in consumer products, or in the environment after disposal. Toxicity is assessed through a battery of tests. Here are some results from a study of the toxicity of black newsprint ink in 7-week-old female rats. The rats' fur was locally clipped twice a week for 13 weeks. One group of rats received a dermal application of ink right after each clipping, and a control group of rats was left untreated. Here are the body weights (in grams) of female rats at the beginning of the study and at the end of the 13 weeks:[24]

Control group		Treatment group	
Week 0	**Week 13**	**Week 0**	**Week 13**
111.2	191.6	107.3	187.0
105.4	191.2	116.7	189.5
110.8	210.7	112.2	179.2
105.6	185.2	103.4	172.2
106.1	195.0	113.2	178.7
104.4	188.3	110.6	180.9
114.0	188.4	110.6	188.3
115.1	195.6	100.5	188.9

109.2	204.6	106.3	183.1
111.3	195.7	112.5	184.5

First, verify that the two experimental groups are not significantly different at the beginning of the study. Then examine the data for evidence that ink application impairs growth in female rats between 7 and 20 weeks of age. That is, do rats treated with ink gain less weight than control rats overall?

27.48 Food safety at restaurants. Example 27.5 describes a study of the attitudes of people attending outdoor fairs about the safety of the food served at such locations. The full data file is available on the companion website as the file *ex27-48.dat.* The variable “srest” contains responses to the same question asked about food served in restaurants. The variable “gender” contains F if the respondent is a woman, M if he is a man. We saw that women are more concerned than men about the safety of food served at fairs. Is this also true for restaurants?

How does the meeting of large rivers influence the diversity of fish? A study of the Amazon and 13 of its major tributaries concentrated on electric fish, which are common in South America. The researchers trawled in more than 1000 locations in the Amazon above and below where each tributary enters the Amazon and in the lower parts of the tributaries themselves. In all, they found 43 species of electric fish. These distinctive fish can “stand in” for fish in general, which are too numerous to count easily. The researchers concluded that the number of fish species increases when a tributary joins the Amazon but that the effect is local: There is no steady increase in diversity as we move downstream. Table 27.3 gives the estimated numbers of electric fish species in the Amazon upstream and downstream from each tributary and in the tributaries themselves just before they flow into the Amazon.[25] *The researchers used nonparametric tests to assess the statistical significance of their results. Exercises 27.49 to 27.51 quote conclusions from the study.*

TABLE 27.3 **Electric fish species in the Amazon**

	Species counts		
Tributary	**Upstream**	**Tributary**	**Downstream**

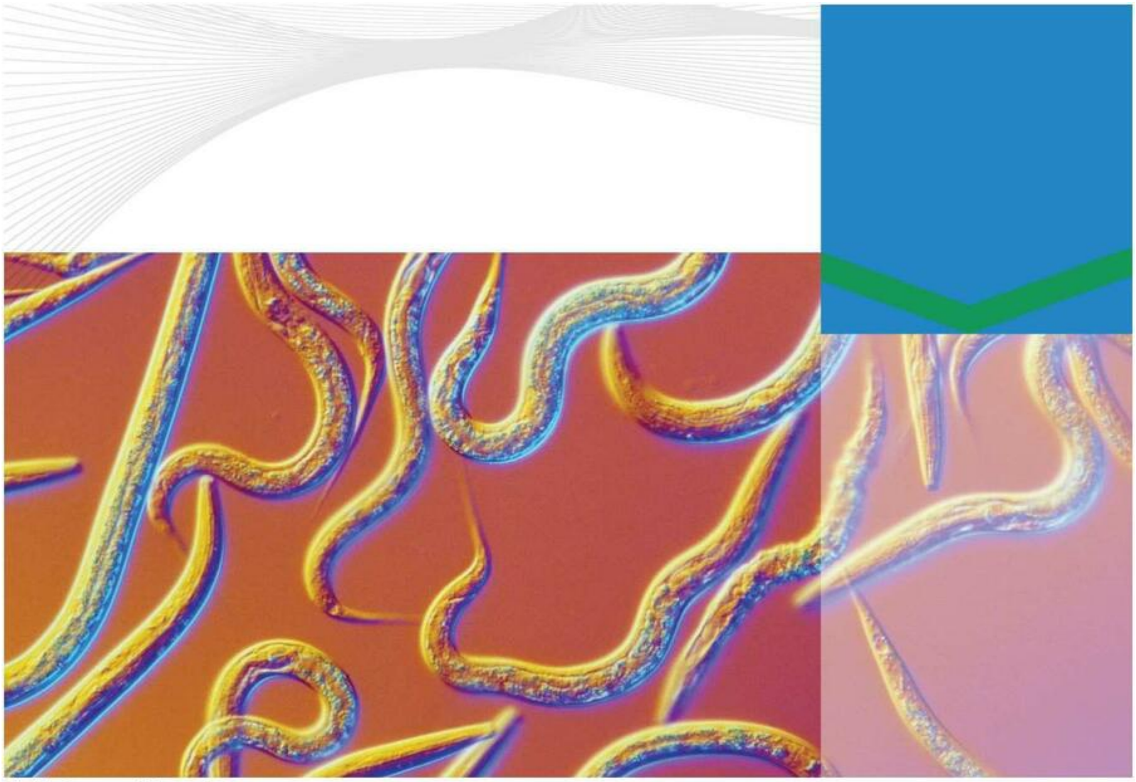
Iça	14	23	19
JutaÍ	11	15	18
Juruá	8	13	8
Japurá	9	16	11
Coari	5	7	7
Purus	10	23	16
Manacapuru	5	8	6
Negro	23	26	24
Madeira	29	24	30
Trombetas	19	20	16
Tapajós	16	5	20
Xingu	25	24	21
Tocantins	10	12	12

27.49 Downstream versus upstream. “We identified a significant positive effect of tributaries on Amazon mainstream species richness in two respects. First, we found that sample stations downstream of each tributary contained more species than did their respective upstream stations.” Do a test to confirm the statistical significance of this effect and report your conclusion.

27.50 Tributary versus upstream. “Second, we found that species richness within tributaries exceeded that within their adjacent upstream mainstream stations.” Again, do a test to confirm significance and report your finding.

27.51 Tributary versus downstream. Species richness “was comparable between tributaries and their adjacent downstream mainstream stations.” Verify this conclusion by comparing tributary and downstream species counts.

CHAPTER 28 Multiple and Logistic Regression



Sinclair Stammers/Science Source

IN THIS CHAPTER WE COVER...

- Parallel regression lines
- Estimating parameters
- Using technology
- Conditions for inference
- Inference for multiple regression
- Interaction
- A case study for multiple regression
- Logistic regression

- Inference for logistic regression

When a scatterplot shows a linear relationship between a quantitative explanatory variable x and a quantitative response variable y, we fit a regression line to the data to describe the relationship. We can also use the line to predict the value of y for a given value of x. For example, Chapter 4 uses regression lines to describe relationships between

- the gain in body fat y of overfed individuals and the amount of nonexercise activity x in their daily routine
- the mating-call frequency y of frogs and outside temperature x
- the number y of new adults that join a colony of birds and the percent x of adult birds that return from the previous year.

In all these cases, other explanatory variables might improve our understanding of the response y and help us to better predict y.

- The gain in body fat y of an overfed individual may depend on the amount of daily nonexercise activity x_1 and on the individual's age x_2 and gender x_3.
- Mating-call frequency y may depend on outside temperature x_1 and also on the species x_2 of the frogs we study.
- The number y of new adults in a bird colony may depend on the percent x_1 of returning adults and also on the species x_2 of the birds we study.

We will now call regression with just one explanatory variable **simple linear regression** to remind us that this is a special case. This chapter introduces the more general case of **multiple regression,** which allows several explanatory variables to combine in explaining a response variable. We start with the simplest case by comparing two parallel regression lines and will gradually build up to increasingly complex models.

PARALLEL REGRESSION LINES

In Chapter 3 we learned how to add a categorical variable to a scatterplot by using different colors or plot symbols to indicate the different values of the categorical variable. In Example 3.7 we examined the relationship between the longevity y and the thorax length of male fruit flies under two experimental conditions. The conditions make up a categorical variable (call it x_2) that takes just two values. The scatterplot shows two *parallel* straight-line patterns linking the response y (longevity) to a quantitative explanatory variable x_1 (thorax length), with one pattern for each value of x_2 (experimental condition).

EXAMPLE 28.1 Cost of reproduction in male fruit flies

STATE: Longevity in male fruit flies is positively associated with adult size. But reproduction has a high physiological cost that could impact longevity. A study looks at the association between longevity and adult size in male fruit flies kept under one of two conditions. One group is kept with sexually active females over the male's life span. The other group is cared for in the same way but kept with females that are not sexually active. Table 28.1 gives the longevity in days and thorax length in millimeters (mm) for the male fruit flies given an opportunity to reproduce (IndReprod = 1) and for those deprived of the opportunity (IndReprod = 0).[1] Is there evidence that reproduction impacts male fruit fly longevity?

TABLE 28.1 Thorax length (mm) and longevity (days) of reproducing and nonreproducing male fruit flies

Longevity	ThxLength	IndReprod	Longevity	ThxLength	IndReprod
35	0.64	0	16	0.64	1
37	0.68	0	19	0.64	1
49	0.68	0	19	0.68	1
46	0.72	0	33	0.72	1
64	0.72	0	34	0.72	1
39	0.76	0	34	0.74	1

46	0.76	0	30	0.76	1
56	0.76	0	42	0.76	1
64	0.76	0	42	0.76	1
65	0.76	0	34	0.78	1
56	0.80	0	26	0.80	1
65	0.80	0	30	0.80	1
70	0.80	0	40	0.82	1
64	0.84	0	54	0.82	1
65	0.84	0	35	0.84	1
70	0.84	0	35	0.84	1
76	0.84	0	46	0.84	1
81	0.84	0	46	0.84	1
85	0.84	0	42	0.88	1
70	0.88	0	46	0.88	1
70	0.88	0	54	0.88	1
76	0.92	0	54	0.88	1
76	0.92	0	56	0.88	1
76	0.94	0	61	0.88	1
			44	0.92	1

PLAN: Make a scatterplot to display the relationship between longevity y and thorax length x_1. Use different colors for the two reproductive conditions. (So reproduction is a categorical variable x_2 that takes two values.) If both reproductive conditions show linear patterns, fit two separate least-squares regression lines to describe them.

SOLVE: Figure 28.1 shows a scatterplot with two separate regression lines, one for each experimental group. The longevities for fruit flies that never reproduced are higher than the longevities of reproducing fruit flies of similar size. Both regression lines have a positive slope, indicating that larger adult males live longer, but the males prevented from reproducing clearly have greater longevity than reproducing males of similar size. Software gives the following regression lines:

For nonreproducing males: y^=-52.381+143.504x1For reproducing males: y^=-59.816+123.370x1

$$\begin{aligned}\text{For nonreproducing males: } \hat{y} &= -52.381 + 143.504x_1 \\ \text{For reproducing males: } \hat{y} &= -59.816 + 123.370x_1\end{aligned}$$

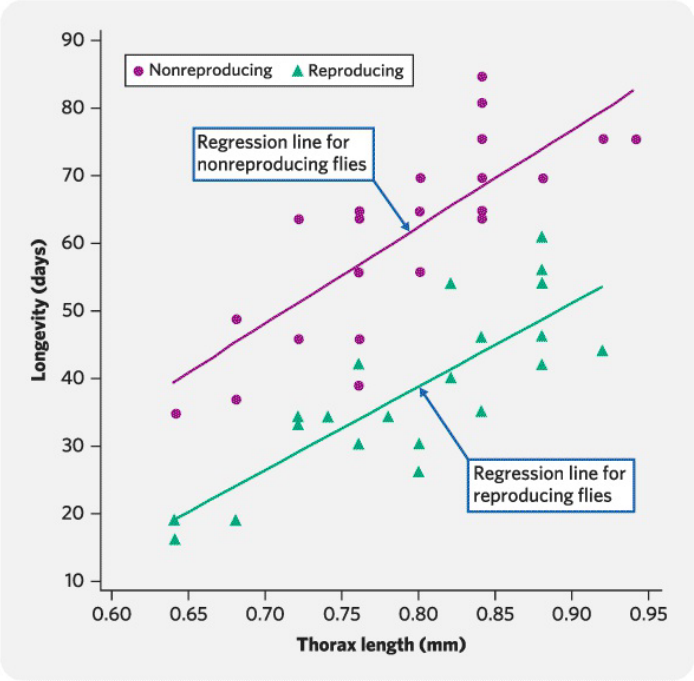


Figure 28.1
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.1 A scatterplot of longevity of male fruit flies of different adult sizes, with two separate regression lines, for Example 28.1.

We notice from the estimated regression lines and Figure 28.1 that the intercepts for the two regression lines are clearly different, but the two regression lines are roughly parallel. Therefore, the difference in the two intercepts indicates how much longer the nonreproducing males live after we take into account thorax length.We will soon learn how to formally estimate parameters and make inferences for parallel regression lines.

CONCLUDE: Our preliminary analysis clearly shows that the nonreproducing male fruit flies live longer than reproducing males of similar size. We know that both groups of male fruit flies lived in similar conditions, except the nonreproducing males lived their lives in the company of females that were not fertile and could not reproduce, whereas the reproducing males lived with fertile females. So the physiological cost of reproducing was the sole difference between the two groups. We will learn later in this chapter how to formally test if the

> difference observed is statistically significant.

We now think that the relationship between longevity and thorax length has about the same rate for both experimental conditions, but that the physiological cost of reproduction results in shorter life spans. This difference in life span appears to be fairly constant for all thorax lengths. We would like to have a single regression model that captures this insight.

To do this, introduce a second explanatory variable x_2 for "reproductive status." This is a categorical variable that takes the values "reproductive" and "nonreproductive." To incorporate this categorical variable x_2 into a regression model, we simply use values 0 and 1 to distinguish the two conditions. Now we have an *indicator variable:*

x2=0 for nonreproductive malesx2=1 for reproductive males

$$\begin{aligned} x_2 &= 0 \text{ for nonreproductive males} \\ x_2 &= 1 \text{ for reproductive males} \end{aligned}$$

INDICATOR VARIABLE

An **indicator variable** places individuals into one of two categories, usually coded by the two values 0 and 1.

An indicator variable is like an indicator light on the dash of a car. If the fuel is above a certain level, the indicator light for fuel is off, but if the fuel drops below a certain level, the light switches on to indicate that there is a low amount of fuel in the tank. An indicator variable for fuel, say x_2, could be coded as $x_2 = 0$ if the fuel is above a certain level and $x_2 = 1$ if the fuel is below that level. Indicator variables are commonly used to indicate categorical characteristics such as gender (0 = male, 1 = female) or condition of patient (0 = good, 1 = poor). When a categorical variable can take more than two values, it is necessary to use more than one indicator variable to encode it. For instance, if we describe a patient's condition as stationary, improving, or deteriorating, we would need two indicator variables, say x_2 and x_3, each taking only values 0 or 1; we could then use stationary = (1, 0), improving = (0, 1), and deteriorating = (0, 0) in the regression model.

The conditions for inference in simple linear regression (Chapter 23, page 577) describe the relationship between the explanatory variable *x* and the mean response μ_y in the population by a *population regression line* $\mu_y = \alpha + \beta x$. In this chapter we switch to the notation $\mu_y = \beta_0 + \beta_1 x$ because it is easier to extend to the multiple regression model. Now we add a second explanatory variable, so that our *regression model* for the population becomes

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The conditions for inference are the same as in the simple linear regression setting: For any fixed values of the explanatory variables, *y* varies about its mean according to a Normal distribution with unknown standard deviation σ, which is the same for all values of x_1 and x_2. We will look in detail at conditions for inference in multiple regression later on.

EXAMPLE 28.2 Interpreting a multiple regression model

Multiple regression models are no longer simple straight lines, so we must think a bit harder in order to interpret what they say. Consider our model

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

in which *y* is the longevity of male fruit flies in the lab, x_1 is their thorax length, and x_2 is an indicator variable for reproductive status. For nonreproducing males, $x_2 = 0$, so the model becomes

$$\mu_y = \beta_0 + \beta_1 x_1$$

For reproducing males, $x_2 = 1$, so the model is

$$\begin{aligned} \mu_y &= \beta_0 + \beta_1 x_1 + \beta_2 \\ &= (\beta_0 + \beta_2) + \beta_1 x_1 \end{aligned}$$

Look carefully: The slope that describes how the mean longevity changes as thorax length x_1 varies is β_1 in both groups. The intercepts differ: β_0 for nonreproducing males and (β0+β2$\beta_0+\beta_2$) for reproducing males. So β_2 is of particular interest, because it is the fixed change between nonreproducing and reproducing male fruit flies.

Figure 28.2 is a graph of this model with all three β’s identified. We have succeeded in giving a single model for two parallel straight lines.

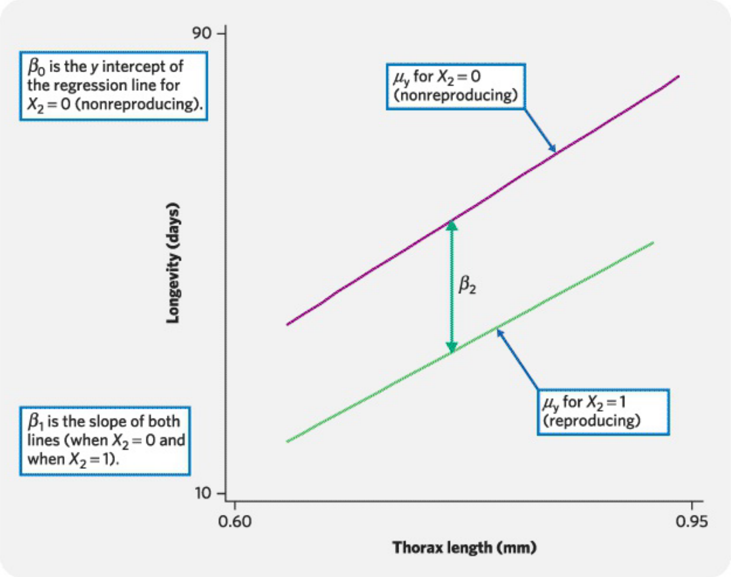


Figure 28.2
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.2 Multiple regression model with two parallel straight lines, for Example 28.2.

APPLY YOUR KNOWLEDGE

28.1 Bird colonies. Suppose that the number y of new birds that join a colony this year has the same straight-line relationship with the percent x_1 of returning birds in colonies for two different bird species. An indicator variable shows which species we observe; $x_2 = 0$ for one, and $x_2 = 1$ for the other. Write a population regression model that describes this setting. Explain in words what each β in your model means.

ESTIMATING PARAMETERS

How shall we estimate the *β*'s in the model μy=β0+β1x1+β2x2? $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$? Because we hope to predict *y*, we want to make the errors in the *y* direction small. We can't call this the vertical distance from the points to *a line* as we did for a simple linear regression model, because we now have two lines. But we still concentrate on the prediction of *y* and therefore on the deviations between the observed responses *y* and the responses predicted by the regression model. These deviations are residual errors, similar to the residuals we first saw in simple linear regression:

residual=observed y-predicted y=y-y^

$$\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$$

The method of least squares estimates the *β*'s in the model by choosing the values that minimize the sum of the squared residuals. Call *b*'s the estimates of the *β*'s. Then the *b*'s minimize

∑(residual)2=∑(observedy-predictedy)2=∑(y-y^)2=∑(y-b0-b1x1-b2x2)2

$$\begin{aligned}\sum(\text{residual})^2 &= \sum(\text{observed}y - \text{predicted}y)^2 \\ &= \sum(y-\hat{y})^2 \\ &= \sum(y - b_0 - b_1x_1 - b_2x_2)^2\end{aligned}$$

The least-squares regression model y^=b0+b1x1+b2x2$\hat{y} = b_0 + b_1x_1 + b_2x_2$ estimates the population regression model μy=β0+β1x1+β2x2. $\mu_y = \beta_0 + \beta_1x_1 + \beta_2x_2$. Remember from Chapter 23 that the population model describes the mean response μ_y for given values of the explanatory variable(s).

The remaining parameter to be estimated is the standard deviation *σ*, which describes the variability of the response *y* about the mean μ_y given by the population regression model. Since the residuals estimate how much *y* varies about the mean of the regression model, the standard deviation *s* of the residuals is used to estimate *σ*. The value of *s* is also referred to as the *regression standard error*.

REGRESSION STANDARD ERROR

The **regression standard error** for the multiple regression **model with two parallel lines** is s=1n-3 ∑residual2=1n-3 ∑(y-y^)2

$$s=\sqrt{\frac{1}{n-3}\sum\text{residual}^2}=\sqrt{\frac{1}{n-3}\sum(y-\hat{y})^2}$$

where *n* is the number of observations in the data set. Use *s* to estimate the standard deviation *σ* of the responses about the mean given by the population regression model.

Notice that instead of dividing by (n-2$n-2$), as we did for the simple linear regression model in Chapter 23, we are now dividing by (n-3$n-3$). Since we are estimating three *β* parameters in our population regression model, the degrees of freedom must reflect this change. In general, the **degrees of freedom** for the regression standard error will be the number of data points minus the number of *β* parameters in the population regression model.

Why do we prefer one regression model with parallel lines to the two separate regression models? Simplicity is one reason—why use separate models with four *β*’s if a single model with three *β*’s describes the data well? Looking at the regression standard error provides another reason: The *n* in the formula for *s* includes all the observations in both groups. As usual, more observations produce a more precise estimate of *σ*. (Of course, using one model for both groups assumes that *σ* describes the scatter about the line in both groups.)

EXAMPLE 28.3 Cost of reproduction in male fruit flies

We developed a single model with parallel straight lines in Example 28.2 for the longevities in the two experimental groups shown in Figure 28.1. The mean longevity is μy=β0+β1x1+β2x2,$\mu_y=\beta_0+\beta_1x_1+\beta_2x_2$, where x_1 is the individual thorax length (the value on the *x* axis in Figure 28.1) and x_2 is an indicator variable to identify the experimental group (different symbols in Figure 28.1). The estimated regression model obtained from statistical software is

y^=-44.285+133.395x1-23.551x2

$$\hat{y}=-44.285+133.395x_1-23.551x_2$$

By substituting the two values of the indicator variable into our estimated regression equation, we can obtain a least-squares line for each group. The predicted longevities are

y^=-67.836+133.395x1 for reproducing males (x2=1)

$$\hat{y} = -67.836 + 133.395x_1 \text{ for reproducing males } (x_2 = 1)$$

and

y^=-44.285+133.395x1 for nonreproducing males (x2=0)

$$\hat{y} = -44.285 + 133.395x_1 \text{ for nonreproducing males } (x_2 = 0)$$

Comparing these estimated regression equations with the two separate regression lines obtained in Example 28.1, we see that the intercept parameters are very close to one another (−67.836 is close to −59.816, and −44.285 is close to −52.381) for both groups. The big change, as intended, is that the slope 133.395 is now the same for both lines. In other words, the estimated change in mean longevity for a 1-unit change in thorax length is now the same for both models, 133.395. A closer look reveals that 133.395 is roughly the average of the two slope estimates (143.504 and 123.370) obtained in Example 28.1.

Finally, the regression standard error s=7.846$s = 7.846$ indicates the size of the error between the observations and the estimated model.

APPLY YOUR KNOWLEDGE

28.2 LDL cholesterol. A study examined the change in blood LDL cholesterol level (in millimoles per liter, mmol/l) as a function of baseline LDL cholesterol level when subjects took either a placebo or the drug Pravastatin for two years.[2] One model of the results shows two parallel lines and gives a multiple linear regression model μy=-1.875+0.4375x1+1.4x2,$\mu_y = -1.875 + 0.4375x_1 + 1.4x_2$, where x_1 represents baseline cholesterol level and x_2 is an indicator variable for treatment that takes value 0 for the placebo group and 1 for the Pravastatin group. Identify the parameter estimates in the model. Find the straight-line equations for the placebo group and for the Pravastatin group, and then make a graph that shows these two lines.

USING TECHNOLOGY

Table 28.1 provides a compact way to display data in a textbook, but this is typically not the best way to enter your data into a statistical software package for analysis. The primary disadvantage to entering the data into a worksheet as it appears in Table 28.1 is that the response variable of interest, longevity, is entered in two separate columns, one for each experimental group.

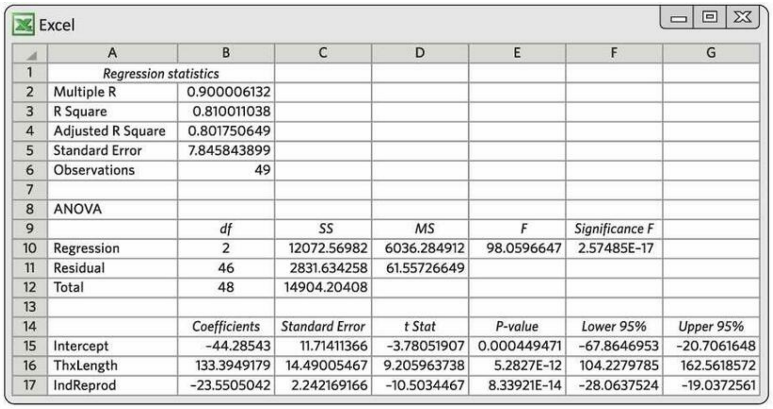
As problems become more complex, we often collect information on more than two variables. This means that data management becomes a much larger part of our job. When fitting and analyzing models, it is usually best to have one column for each variable and to use the rows to identify the individuals.

EXAMPLE 28.4 Organizing data

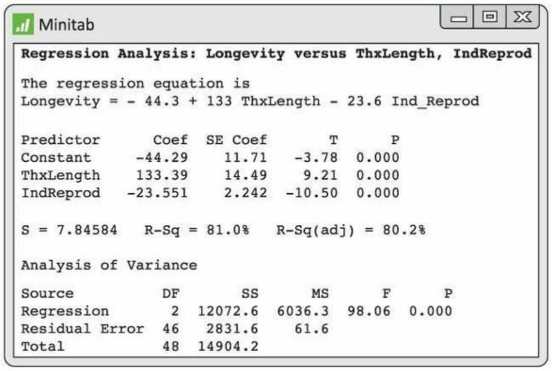
To fit the multiple regression model with equal slopes in Example 28.3, three columns were created. The 49 longevities y for reproducing and nonreproducing males were stacked into a column labeled *Longevity*, values of the explanatory variable x_1 were entered into a column labeled *ThxLength*, and values of the indicator variable x_2 were entered into a column labeled *IndReprod*. Here are the first five rows of the worksheet:

Row	Longevity	ThxLength	IndReprod
1	35	0.64	0
2	37	0.68	0
3	49	0.68	0
4	46	0.72	0
5	64	0.72	0

Statistical software performs all the necessary computations after we identify the response variable *Longevity* and the two explanatory variables *ThxLength* and *IndReprod*. Figure 28.3 shows the regression output from Excel, Minitab, R, and SPSS. The format of the output differs slightly, but each package provides parameter estimates, standard errors, *t* statistics, and *P*-values, an analysis of variance (ANOVA) table, the regression standard error, and R^2. We will digest this output one piece at a time: first describing the model, then looking at the conditions needed for inference, and finally interpreting the results of inference.

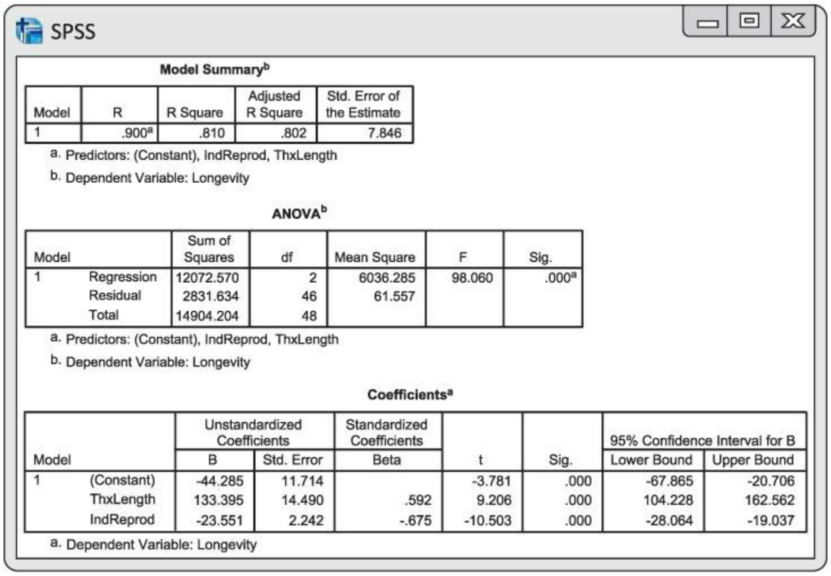


	A	B	C	D	E	F	G
1	*Regression statistics*						
2	Multiple R	0.900006132					
3	R Square	0.810011038					
4	Adjusted R Square	0.801750649					
5	Standard Error	7.845843899					
6	Observations	49					
7							
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*Significance F*	
10	Regression	2	12072.56982	6036.284912	98.0596647	2.57485E-17	
11	Residual	46	2831.634258	61.55726649			
12	Total	48	14904.20408				
13							
14		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
15	Intercept	-44.28543	11.71411366	-3.78051907	0.000449471	-67.8646953	-20.7061648
16	ThxLength	133.3949179	14.49005467	9.205963738	5.2827E-12	104.2279785	162.5618572
17	IndReprod	-23.5505042	2.242169166	-10.5034467	8.33921E-14	-28.0637524	-19.0372561



```
Regression Analysis: Longevity versus ThxLength, IndReprod

The regression equation is
Longevity = - 44.3 + 133 ThxLength - 23.6 Ind_Reprod


Predictor       Coef  SE Coef       T      P
Constant      -44.29    11.71   -3.78  0.000
ThxLength     133.39    14.49    9.21  0.000
IndReprod    -23.551    2.242  -10.50  0.000

S = 7.84584   R-Sq = 81.0%   R-Sq(adj) = 80.2%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2  12072.6  6036.3  98.06  0.000
Residual Error  46   2831.6    61.6
Total           48  14904.2
```

Figure 28.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.900[a]	.810	.802	7.846

a. Predictors: (Constant), IndReprod, ThxLength
b. Dependent Variable: Longevity

ANOVA[b]

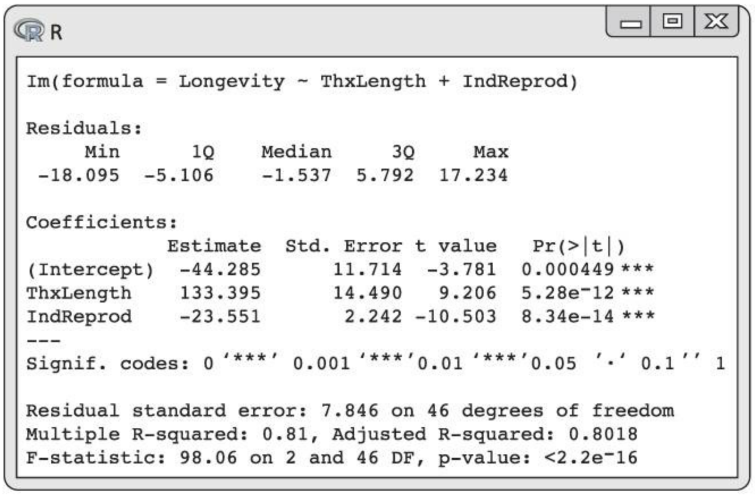
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12072.570	2	6036.285	98.060	.000[a]
	Residual	2831.634	46	61.557		
	Total	14904.204	48			

a. Predictors: (Constant), IndReprod, ThxLength
b. Dependent Variable: Longevity

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-44.285	11.714		-3.781	.000	-67.865	-20.706
	ThxLength	133.395	14.490	.592	9.206	.000	104.228	162.562
	IndReprod	-23.551	2.242	-.675	-10.503	.000	-28.064	-19.037

a. Dependent Variable: Longevity



```
Im(formula = Longevity ~ ThxLength + IndReprod)

Residuals:
     Min       1Q    Median     3Q     Max
 -18.095  -5.106    -1.537  5.792  17.234

Coefficients:
             Estimate  Std. Error t value   Pr(>|t|)
(Intercept)  -44.285       11.714  -3.781  0.000449 ***
ThxLength    133.395       14.490   9.206  5.28e-12 ***
IndReprod    -23.551        2.242 -10.503  8.34e-14 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘***’0.01 ‘***’0.05  ’·’ 0.1’’ 1

Residual standard error: 7.846 on 46 degrees of freedom
Multiple R-squared: 0.81, Adjusted R-squared: 0.8018
F-statistic: 98.06 on 2 and 46 DF, p-value: <2.2e-16
```

Figure 28.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.3 Output from Excel, Minitab, SPSS, and R for the model with parallel regression lines in Example 28.3.

EXAMPLE 28.5 Parameter estimates on statistical output

On the Excel output in Figure 28.3, the parameter estimates b0=-44.28543,$b_0 = -44.28543$, b1=133.3949179,$b_1 = 133.3949179$, and b2=-23.5505042$b_2 = -23.5505042$ are clearly labeled in a column. Thus, our multiple regression model for predicting longevity (after rounding) is y^=-44.285+133.395x1-23.551x2.$\hat{y} = -44.285 + 133.395x_1 - 23.551x_2$. SPSS and R have a similar output format. Minitab provides the estimated regression equation first and then gives more detailed estimates in a column labeled “Coef” for coefficients. After rounding, all four sets of estimates match the estimates provided in Example 28.3.

Although the labels differ again, the regression standard error is provided by all four packages:

Excel:Standard Error=7.845843899Minitab:S=7.84584R:Residual Standard Error=7.846SPSS:Std. Error of the Estimate=7.846

$$\begin{array}{ll} \text{Excel :} & \text{Standard Error} = 7.845843899 \\ \text{Minitab :} & S = 7.84584 \\ \text{R :} & \text{Residual Standard Error} = 7.846 \\ \text{SPSS :} & \text{Std. Error of the Estimate} = 7.846 \end{array}$$

For simple linear regression models, the square of the correlation coefficient r^2 between y and x provides a statistic that can be used along with residual plots and other techniques to assess the fit of the model. In particular, r^2 measures the proportion of variation in the response variable that is explained by using the explanatory variable. For our multiple regression model with parallel regression lines, we do not have one correlation coefficient. However, by squaring the correlation coefficient between the observed responses y and the predicted responses y^,$\hat{y}$, we obtain the *squared multiple correlation coefficient* R^2.

Alternative computation formulas based on values in the ANOVA table of the output help us interpret this new statistic. You may want to refer to Chapter 24 to refresh your memory about the details of the ANOVA procedure. The ANOVA table breaks the total variability in the responses into two pieces. One piece summarizes the variation explained by the model, and the other piece summarizes the residual variation traditionally labeled “error.” In short, we have

total variation=variation explained by model+residual variation

$$\text{total variation} = \text{variation explained by model} + \text{residual variation}$$

The value of R^2 is the ratio of *model* to *total* variation, so R^2 tells us how much variation in the response variable y we explained by using the set of explanatory variables in the multiple regression model.

SQUARED MULTIPLE CORRELATION COEFFICIENT

The **squared multiple correlation coefficient** R^2 is the square of the correlation coefficient between the observed responses *y* and the predicted responses y^$\hat{y}$ and can be computed as

R2=variability explained by modeltotal variability in y=∑(y^-y-)2∑(y-y-)2=SSModelSSTotal

$$R^2 = \frac{\text{variability explained by model}}{\text{total variability in} y} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{\text{SSModel}}{\text{SSTotal}}$$

The denominator measures the deviation of the observed responses about their mean. Just as in simple linear regression, the predicted responses y^$\hat{y}$ have the same mean $\bar{y}$ as the observed responses. So the numerator is the variability we would see if the model fitted perfectly and there were no spread of the *y*'s about the model. We can think of this as the variability explained by the model.

EXAMPLE 28.6 Using R^2

The value of R2=0.810$R^2 = 0.810$ for our multiple regression model with parallel lines in Example 28.3 can be found in the output for all four packages in Figure 28.3 (labeled as Multiple R-squared, R Square, or R-Sq). Thus, the proportion of variation in the response variable *Longevity* that is explained by the explanatory variable *ThxLength* and the indicator variable *IndReprod* using the regression model with parallel lines is 0.810, or 81.0%. Converting R^2 to the percent of variation explained by the explanatory variables in a multiple regression model is a common practice. The value of R^2 indicates that 81% of the variation in the longevity of male fruit flies in the lab is explained by using multiple regression with an explanatory variable for their thorax length and an indicator variable for the experimental condition determining their reproductive status.

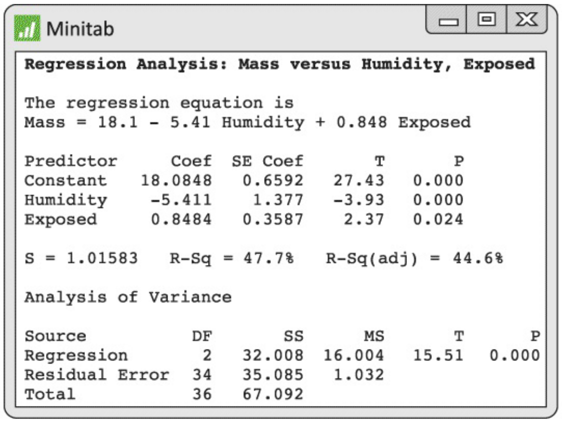
The squared multiple correlation coefficient takes only values between 0 and 1 (0% and 100%) and is a very useful statistic to help us assess the fit of a multiple regression model.

APPLY YOUR KNOWLEDGE

28.3 Heights and weights for boys and girls. Suppose you are designing a study to investigate the relationship between height and weight for boys and girls.

a. Specify a model with two regression lines that could be used to predict height separately for boys and for girls. Be sure to identify all variables and describe all parameters in your model.
b. How many columns in a worksheet would be required to fit this model with statistical software? Describe each column.

28.4 Nestling mass and nest humidity. Researchers investigated the relationship between nestling mass, measured in grams, and nest humidity index, measured as the ratio of total mass of water in the nest divided by nest dry mass, for two different groups of great titmice parents.[3] One group was exposed to fleas during egg laying, and the other was not. Exposed parents were coded as 1, and unexposed parents were coded as 0. Use the output provided, obtained by fitting a multiple regression model with parallel lines for the two groups of parents, to answer the following questions.



```
Regression Analysis: Mass versus Humidity, Exposed

The regression equation is
Mass = 18.1 - 5.41 Humidity + 0.848 Exposed

Predictor      Coef  SE Coef       T      P
Constant    18.0848   0.6592   27.43  0.000
Humidity     -5.411    1.377   -3.93  0.000
Exposed      0.8484   0.3587    2.37  0.024

S = 1.01583   R-Sq = 47.7%   R-Sq(adj) = 44.6%

Analysis of Variance

Source          DF       SS      MS       T      P
Regression       2   32.008  16.004   15.51  0.000
Residual Error  34   35.085   1.032
Total           36   67.092
```

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

a. Identify the regression model for predicting nestling mass from nest humidity index for the two groups of great titmice parents.

b. Based on your model, do you think that nestling mass was higher in nests of birds exposed to fleas during egg laying? Explain.
c. What is the value of the regression standard error? Interpret this value.
d. What is the value of the squared multiple correlation coefficient? Interpret this value.

CONDITIONS FOR INFERENCE

We have seen in a simple but useful case how adding another explanatory variable can fit patterns more complex than the single straight line of simple linear regression. So far we have included two explanatory variables: a quantitative variable x_1 and an indicator variable x_2. Multiple linear regression, however, can allow any number of explanatory variables, each of which can be either quantitative or indicator. You can also work with data from just one random sample, or you can pull together separate random samples represented by an indicator variable (as in the case of the fruit fly data in Example 28.1). Here is a statement of the general model that includes the conditions needed for inference.

THE MULTIPLE LINEAR REGRESSION MODEL

We have *n* observations on *k* explanatory variables x1,x2,…,xk $x_1, x_2, \ldots, x_k$ and a response variable *y*. Our goal is to study or predict the behavior of *y* for a given set of the explanatory variables.

- For any set of fixed values of the explanatory variables, the response *y* varies according to a **Normal distribution.** Repeated responses *y* are **independent** of each other.
- The mean response μ_y has a **linear relationship** given by the **population regression model**

$$μy=β0+β1x1+β2x2+⋯+βkxk$$

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

The *β*’s are unknown parameters.

- The **standard deviation** of *y* (call it *σ*) is the same for all values of the explanatory variables. The value of *σ* is unknown.

This model has k+2$k+2$ parameters that we must estimate from data: the k+1$k+1$ coefficients β_0, β1,…,βk$\beta_1, \ldots, \beta_k$ and the standard deviation *σ*.

These conditions are similar to those we have described in Chapter 23 for simple linear regression (page 577). However, the challenge in checking inference conditions for multiple regression is that there is no one simple plot that can give the entire picture. We will present a few simple basics here because regression diagnostics is a subject that could be expanded to several chapters.

First, plot the response variable against each of the explanatory variables. These plots help you explore and understand potential relationships. Multiple regression models allow curvature and other interesting features that are not simple to visually check, especially when we get beyond two explanatory variables. For example, if you see a quadratic pattern, you should consider adding a quadratic term (x^2) for that explanatory variable.

Then, plot the residuals against the predicted values and all of the explanatory variables in the model. These plots will allow you to check the condition that **the standard deviation of the response about the multiple regression model is the same everywhere.** They should show an unstructured horizontal band of points centered at 0. The mean of the residuals is always 0, just as in simple linear regression. Funnel or cone shapes indicate that this condition is not met and that one or more variables should be transformed to stabilize the standard deviation of the residuals before doing inference.

Look for outliers and influential observations in all residual plots. To check how much influence a particular observation has, you can fit your model with and without this observation. If the estimates and statistics do not change much, you can safely proceed. However, if there are substantial changes, you must begin a more careful investigation. Do notsimply throw out observations to improve the fit and increase R^2. (See Chapter 2, page 55, for an in-depth discussion of outlier treatment.)

Ideally, we would like all the explanatory variables to be independent and the observations on the response variable to be independent. As you will see in this chapter, practical problems include explanatory variables that are not independent. Association between two or more explanatory variables can create serious problems in the model, so use correlations and scatterplots to check relationships.

The response should vary about the multiple regression model

according to a Normal distribution. This condition is checked by making a histogram, stemplot, or Normal quantile plot of the residuals. Once again, we will rely on the robustness of the regression methods when there is a slight departure from Normality, except for prediction intervals. As we did with simple linear regression, we view prediction intervals from multiple regression models as rough approximations.

EXAMPLE 28.7 Checking the conditions

Four plots displaying the residuals for the multiple regression model with parallel lines in Example 28.3 are shown in Figure 28.4. The conditions for inference are linearity, Normality, constant variance, and independence. We will check these conditions one at a time.

Linear trend: The scatterplot in Figure 28.1 shows parallel linear patterns for the two experimental groups, so the model is reasonable.

Normality: The normal quantile plot (upper left) and histogram (lower left) of the model's residuals in Figure 28.4 indicate that the residuals are symmetric about zero and approximately Normal.

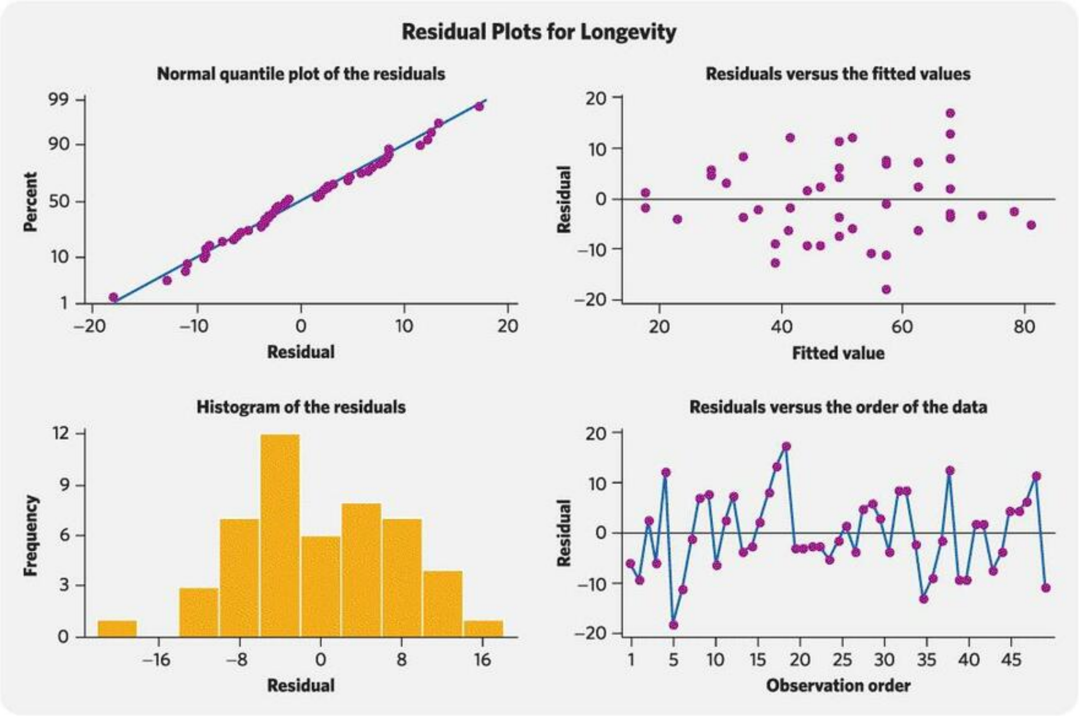


Figure 28.4
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.4 Residual plots to check the conditions for inference for the model with parallel regression lines in Example 28.3.

Constant variance: The residual plot (upper right, residuals versus y^ $\hat{y}$ values) in Figure 28.4 does not have any obvious pattern and shows a fairly constant amount of variability. Note, though, that since there are fewer points at the edges of the plot, it is harder to assess variability for the fruit flies with the shortest and the longest life spans. Overall, this residual plot supports the model's condition that a single σ describes the scatter about the reproducing and nonreproducing lines.

Independence: The male fruit flies were randomly assigned to the two experimental conditions. The individuals should therefore be independent. In addition, the plot of ordered residuals (lower right, ordered as they appear in the data set) provides a quick check to see if there is a pattern in the residuals based on the order in which they were entered into the worksheet. The lack of any obvious pattern in this last graph reinforces our belief that the individuals used for this model are independent.

We conclude that a linear model provides a reasonable summary of the male fruit fly longevities and that the conditions for inference appear to be satisfied. We can now proceed to inference for this model.

APPLY YOUR KNOWLEDGE

28.5 Nutrition and reproduction. A research team looked at the relationship between protein intake (in the form of yeast supplements) and reproductive output (mean number of eggs per female) in wild-type versus mutant fruit flies with higher longevity.[4] Figure 28.5 shows the scatterplot and residual plot of this relationship. Do the plots suggest any potential problems with the conditions for inference? Explain your answer.

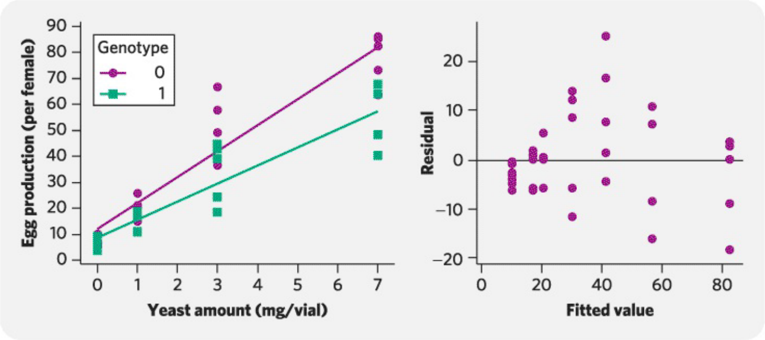
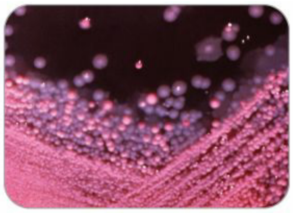


Figure 28.5
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.5 Scatterplot and residual plot to check the conditions for inference in Exercise 28.5.

28.6 Evolution in bacteria. Biologists designed an experiment to test the theory of evolution in *E. coli* bacteria. These bacteria can be grown for 2000 generations ("evolved") in less than a year. Bacteria were evolved in either a neutral or an acidic pH (x_1) and then compared with the ancestral line to compute a relative fitness score (y) in environments of various pH (x_2).[5] A multiple regression model with three variables was found significant with R2=74.1%$R^2 = 74.1\%$ and equation y^=2.12-0.957x1-0.166x2+0.147x1x2. $\hat{y} = 2.12 - 0.957x_1 - 0.166x_2 + 0.147x_1x_2$. The residual plot and a histogram of the residuals are shown in Figure 28.6. Do you see any reason for concern in making inferences from this model, which has a high R^2-value? Comment on both plots.



CDC/Dr. Gilda Jones

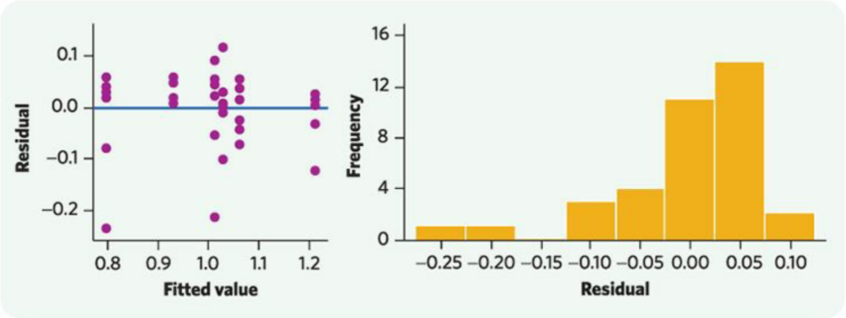


Figure 28.6

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.6 Residual plot and histogram of the residuals to check the conditions for inference in Exercise 28.6.

INFERENCE FOR MULTIPLE REGRESSION

To this point, we have concentrated on understanding the model, estimating parameters, and verifying the conditions for inference that are part of a regression model. Inference in multiple regression begins with tests that help us to decide if a model μy=β0+β1x1+β2x2+⋯+βkxk $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ adequately fits the data and to choose between several possible models.

The first inference for a multiple regression model examines the overall model. The ANOVA table summarizes the breakdown of the variability in the response variable when *k* explanatory variables are used. There is one row for each of the three sources of variation: model, error, and total. Each source of variation has a degrees of freedom associated with it. The *F* statistic for the overall model is the ratio of the model variation (MSM) over the error variation (MSE). Refer back to Chapter 24 on ANOVA if you need to refresh your memory.

F STATISTIC FOR REGRESSION MODEL

The **ANOVA *F*** statistic for testing the null hypotheses that all the regression coefficients (β's) except β_0 are equal to zero has the form

F=variation due to modelvariation due to error=MSMMSE

$$F = \frac{\text{variation due to model}}{\text{variation due to error}} = \frac{\text{MSM}}{\text{MSE}}$$

This *F* statistic is used to compute the *P*-value and determine whether all the *k* regression coefficients (one for each of the *k* explanatory variables, but not the intercept) are significantly different from zero. The *P*-value is obtained by finding the area to the right of *F* under the F(k,n-k-1) $F\,(k, n-k-1)$ distribution. Software typically provides the ANOVA table, which contains all the information you need.

EXAMPLE 28.8 Overall *F* test for parallel lines

The mean longevity is $\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, where x_1 is labeled as *ThxLength* and x_2 is labeled as *IndReprod* on the output in Figure 28.3. The null and alternative hypotheses for the overall *F* test are

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = 0 \text{ (that is, } \mu_y = \beta_0\text{)} \\ H_a &: \text{at least one of } \beta_1 \text{ and } \beta_2 \text{ is not } 0 \end{aligned}$$

The null hypothesis H_0 specifies a model, called the **null model,** where the response variable *y* is a constant (its mean) plus random variation. In other words, the null model says that x_1 and x_2 together do not help predict *y*.

The value of the *F* statistic reported in all four ANOVA tables in Figure 28.3 is $F = 98.06$ (up to rounding error). The *P*-value is obtained from an *F* distribution with 2 degrees of freedom for the numerator and 46 degrees of freedom for the denominator. SPSS and Minitab report a *P*-value of 0.000, which does not mean that its value is zero but that it is less than 0.0005. R shows that $P < 2.2$e-16, and Excel reports that the *P*-value is actually 2.57485E-17, which is extremely small. All four software packages indicate that the *P*-value is extremely significant. Since the *P*-value is less than any reasonable significance level, say, $\alpha = 0.01$, we reject the null hypothesis and conclude that at least one of the *x*'s helps explain the variation in the longevities *y*.

Rejecting the null hypothesis with the *F* statistic tells us that at least one of our β parameters (excluding β_0) is not equal to zero, but it doesn't tell us which parameters are not equal to zero. We turn to individual tests for each parameter to answer that question.

INDIVIDUAL *t* TESTS FOR COEFFICIENTS

To test the null hypothesis that one of the β's in a specific regression model is zero, compute the *t* statistic

$$t = \frac{\text{parameter estimate}}{\text{standard error of estimate}} = \frac{b}{SE_b}$$

$$t = \frac{\text{parameter estimate}}{\text{standard error of estimate}} = \frac{b}{\mathrm{SE}_b}$$

When the conditions for inference are met, the *t* distribution with (n-k-1$n-k-1$) degrees of freedom can be used to compute confidence intervals and conduct hypothesis tests for each of the *β*’s.

EXAMPLE 28.9 Individual *t* tests

The output in Figure 28.3 provides parameter estimates and standard errors for the coefficients β0,β1,β_0, β_1, and β_2. The individual *t* statistic for x_1 (*ThxLength*) tests the hypotheses

H0:β1=0 in this model (that is, μy=β0+β2x2)Ha:β1≠0 in this model

$$H_0 : \beta_1 = 0 \text{ in this model (that is, } \mu_y = \beta_0 + \beta_2 x_2)$$
$$H_a : \beta_1 \neq 0 \text{ in this model}$$

We must state the model explicitly in the null hypothesis because the bare statement H0:β1=0$H_0 : \beta_1 = 0$ can be misleading. The hypothesis of interest is that *in this model* the coefficient of x_1 is 0. If the same x_1 is used in a different model with different explanatory variables, the hypothesis H0:β1=0$H_0 : \beta_1 = 0$ has a different meaning even though we would write it the same way. More on this later.

Using the Minitab output, we see that the test statistic is

t=133.3914.49=9.21 for β1

$$t = \frac{133.39}{14.49} = 9.21 \quad \text{for } \beta_1$$

The *P*-value is obtained by finding the area under a *t* distribution with 49-2-1=46$49 - 2 - 1 = 46$ degreesof freedom below –9.21 or above 9.21. Since this value is so small, Minitab simply reportsthe *P*-value as being 0.000 (<0.0005< 0.0005).

The test statistic for the other variable coefficient is

t=-23.5512.242=-10.50 for β2

$$t = \frac{-23.551}{2.242} = -10.50 \quad \text{for } \beta_2$$

The *P*-value is again obtained using the *t* distribution with 46 degrees of freedom. It is so small that it is reported by Minitab as being 0.000 (<0.0005< 0.0005).

We have very strong evidence that thorax length x_1 (*ThxLength*) helps explain the longevity y even after we allow the reproductive status x_2 (*IndReprod*) to explain longevity. Similarly, reproductive status x_2 adds to our ability to explain longevity even after we take thorax length x_1 into account.

Statistical software reports two-sided *P*-values for the individual *t* tests. If you had an alternative hypothesis that would require a one-sided *P*-value, you can simply divide the *P*-value on the output by 2 if the effect is in the direction of the alternative.

Example 28.9 illustrates one of the easiest situations you will encounter. The overall *F* test tells us that at least one of the coefficients (except for β_0) is not equal to zero; that is, both explanatory variables together help explain the response. Then the individual *t* tests indicate that both coefficients are significantly different from zero. This means that each explanatory variable significantly improves the explanation when added to a model that uses only the other explanatory variable.

Individual *t* tests are helpful in identifying the explanatory variables that are useful predictors, but *extreme caution* is necessary when interpreting the results of these tests. Remember that an individual *t* assesses the contribution of its variable in the presence of the other variables in this specific model. That is, individual *t*'s depend on the model in use, not just on the direct association between an explanatory variable and the response. The case study described later in this chapter illustrates some of these challenges.

CONFIDENCE INTERVALS FOR COEFFICIENTS

A **level C confidence interval for β** is b±t*SEb.$b \pm t^{*}\mathrm{SE}_b$.

The critical value t^* is obtained from the t(n-k-1)$t\,(n-k-1)$ density curve, where *k* is the number of explanatory variables used in the multiple regression model.

EXAMPLE 28.10 Parameter confidence intervals when all predictors are helpful

The individual *t* statistics and the corresponding *P*-values in Figure 28.3 indicate that both explanatory variables are useful predictors. Only Excel and SPSS provide a confidence interval for the regression coefficients. With 95% confidence, both programs determine that β_1 is between 104.228 and 162.562 days per millimeter (mm) of thorax length and that β_2 is within the interval −28.064 to −19.037. Interpretation is relatively simple here because the model has only two parallel lines. We are 95% confident that, in this particular model, the longevity of male fruit flies in each group increases by about 104 to 163 days per mm of thorax length (or, more practically, by 10 to 16 days per tenth of mm) and that reproducing males live about 19 to 28 days less than nonreproducing males of similar size.

If the software you use does not provide confidence intervals for the regression coefficients, you can compute them yourself from the parameter estimates provided. The regression model for the cost of reproduction in male fruit flies has n-k-1=49-2-1=46 $n-k-1=49-2-1=46$ degrees of freedom. Using Table C in the back of the textbook, we find that the *t* critical value for degrees of freedom 40 (since 46 is not in the table) and confidence level 95% is t*=2.021. $t^* = 2.021$. Thus, an approximate 95% confidence interval for β_1 is

b1±t*SEb1=133.395±(2.021)(14.490)=133.395±29.284,or 104.111 to 162.679

$$b_1 \pm t^* \mathrm{SE}_{b1} = 133.395 \pm (2.021)(14.490) = 133.395 \pm 29.284, \text{ or } 104.111 \text{ to } 162.679$$

and an approximate 95% confidence interval for β_2 is

b2±t*SEb2=–23.551±(2.021)(2.242)=–23.551±4.531,or –28.082 to –19.020

$$b_2 \pm t^* \mathrm{SE}_{b2} = -23.551 \pm (2.021)(2.242) = -23.551 \pm 4.531, \text{ or } -28.082 \text{ to } -19.020$$

These approximate intervals are very close to those provided by software.

Interpreting coefficients, beyond significance, can get very challenging in more complex models. We will see an example later in the chapter.

Beyond model building, prediction can also be an important objective of scientific studies. Construction of *confidence intervals for a mean response* and *prediction intervals for a future observation* with multiple regression models is similar to the methods we used for simple linear regression in Chapter 23. The main difference is that we must now specify a list of values for all the explanatory variables in the model. As we learned in simple linear regression, the additional uncertainty in predicting future observations will result in prediction intervals that are wider than confidence intervals.

CONFIDENCE AND PREDICTION INTERVALS FOR MULTIPLE REGRESSION RESPONSE

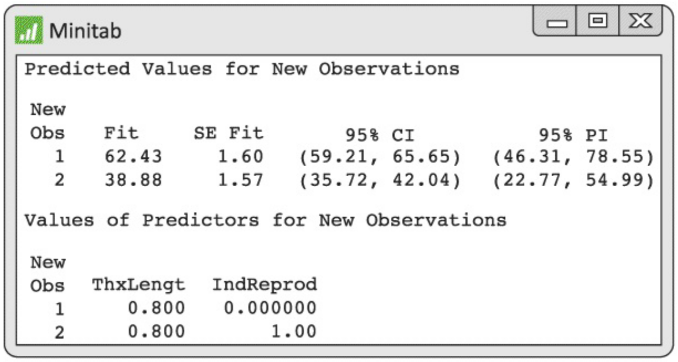
A level *C* **confidence interval for the mean response** μ_y is $\hat{y} \pm t^* \mathrm{SE}_{\hat{\mu}}$.

A level *C* **prediction interval for a single response** ***y*** is $\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$.

The critical value t^* is obtained from the $t(n - k - 1)$ density curve, where *k* is the number of explanatory variables used in the multiple regression model.

EXAMPLE 28.11 Predicting means and future values

Figure 28.7 provides the predicted values (“Fit”), 95% confidence limits for the mean longevity (“95% CI”), and 95% prediction limits (“95% PI”) for two observations. As you can see, we had to specify a numerical value for both of the model’s explanatory variables in order to get these intervals. Here we have obtained confidence and prediction intervals for male fruit flies with thorax lengths of 0.8 mm (whether allowed to reproduce or not). As expected, the prediction limits for future fruit fly longevities are wider than the corresponding confidence limits for the mean fruit fly longevities.



```
Predicted Values for New Observations

 New
 Obs    Fit     SE Fit        95% CI             95% PI
   1   62.43     1.60    (59.21, 65.65)    (46.31, 78.55)
   2   38.88     1.57    (35.72, 42.04)    (22.77, 54.99)

Values of Predictors for New Observations

 New
 Obs   ThxLengt   IndReprod
   1      0.800    0.000000
   2      0.800        1.00
```

Figure 28.7

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.7 Predicted values, confidence limits for mean longevity, and prediction limits for future longevity for two fruit fly observations, for Example 28.11.

APPLY YOUR KNOWLEDGE

28.7 Bone density in elderly men. Bone fractures are a substantial concern among the elderly. A study examined the impact of a number of variables on bone mineral density (BMD) assessed from the femur in a random sample of 43 elderly men hospitalized for hip fracture. The researchers found that BMD could be predicted from serum albumin level and body weight according to the following multiple regression model:[6]

	b	**SE**	***t***	***P***
Constant	−8.14	1.00	−8.12	<0.001
Albumin	0.098	0.026	3.71	<0.01
Weight	0.042	0.01	3.69	<0.01

$F = 15.19$, ($P < 0.001$), $R^2 = 0.451$

Use the information provided to answer the following questions.

a. What is the multiple linear regression equation for this model? What is the number *k* of explanatory variables involved?

b. What tells you that the overall model is significant?

c. What percent of variation in BMD is explained by the model?

d. What is a 95% confidence interval for the slope parameter for albumin level?

e. What is a 95% confidence interval for the slope parameter for body weight?

28.8 Bone density in elderly women. Men and women age differently. The study described in Exercise 28.7 also examined a random sample of 243 elderly women hospitalized for hip fracture. The researchers found that BMD in elderly women with hip fracture could be modeled from body weight, fracture type (cervical or trochanteric), age, and total lymphocyte count (TLC) according to the following multiple regression model:

	b	**SE**	***t***	***P***
Constant	−5.03	0.644	−7.82	<0.001
Weight	0.055	0.005	10.01	<0.001
Fracture	−0.51	0.113	−4.56	<0.001
Age	0.019	0.006	−3.00	<0.01
TLC	3.1E-04	0.000	2.39	<0.05

$F = 42.2$, ($P < 0.001$), $R^2 = 0.422$

Use the information provided to answer the following questions.

a. What is the multiple linear regression equation for this model? What is the number *k* of explanatory variables involved?

b. Which, if any, of the explanatory variables in the model is an indicator variable?

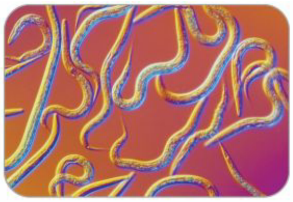
c. What tells you that the overall model is significant?

d. What percent of variation in BMD in elderly women is explained by the model?

e. What is a 95% confidence interval for the slope parameter for body weight?

28.9 Neural mechanism of nociception. The roundworm *Caenorhabditis elegans* is a widely studied animal model, in part because of its small number of neurons and easily manipulated genome. Nociception is the neural perception of an actually or potentially harmful stimulus. In *C. elegans,* it evokes a self-preserving withdrawal behavior. However, repeated stimulation can result in reduced withdrawal response, or habituation. Researchers compared the withdrawal response to disturbing light stimuli in wild-type *C. elegans* and a mutant *C. elegans* line that exhibits a slower response of PVD sensory neurons. Failure to react indicates habituation. Here are the percents of animals tested that exhibited a withdrawal reaction to a noxious stimulus consisting of varying numbers of consecutive light pulses (IndMutant == 1 for the mutant line, and 0 for the wild type):[7]

% Reacting	Pulses	IndMutant	% Reacting	Pulses	IndMutant	% Reacting	Pulses
93	1	1	54	15	1	91	9
81	2	1	56	16	1	82	10
73	3	1	50	17	1	86	11
64	4	1	42	18	1	82	12
73	5	1	42	19	1	91	13
68	6	1	29	20	1	85	14
64	7	1	96	1	0	83	15
65	8	1	100	2	0	70	16
50	9	1	92	3	0	74	17
68	10	1	96	4	0	64	18
57	11	1	92	5	0	68	19
50	12	1	92	6	0	78	20
62	13	1	84	7	0		
44	14	1	96	8	0		



Sinclair Stammers/Science Source

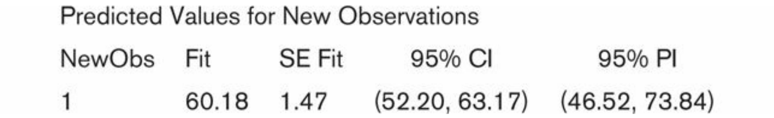
a. Make a scatterplot of the percent reacting as a function of the number of light pulses with two separate lines representing the two subpopulations. Describe the shape of the relationship. What does it suggest about the pattern of withdrawal responses in the two subpopulations of *C. elegans* for increasing numbers of light pulses? How does your answer fit in the context of habituation? Explain why these results suggest an involvement of PVD sensory neurons in nociception and habituation.

b. Use software to obtain the statistical analysis of the linear regression of the percent reacting (y) as a function of the number of light pulses (x_1) and the indicator variable IndMutant (x_2). What is the regression equation for this model?

c. Is the overall model significant? What percent of variation in percent reacting is explained by the model?

d. Are the regression coefficients for the two explanatory variables significant?

28.10 Neural mechanism of nociception: prediction. Minitab gives the following information for mutants exposed to 10 light pulses (x1=10,x2=1$x_1 = 10, x_2 = 1$):



```
Predicted Values for New Observations
NewObs   Fit     SE Fit    95% CI          95% PI
1        60.18   1.47      (52.20, 63.17)  (46.52, 73.84)
```

With 95% confidence, what is the average percent of mutants reacting to 10 light pulses in similar experimental conditions? What range of percents reacting would we predict for 95% of experiments in which mutant *C. elegans* are exposed to 10 light pulses?

INTERACTION

We have first examined simple examples with two parallel linear patterns for two values of an indicator variable. We now turn to situations with two linear patterns that are not parallel. To write a regression model for this setting, we need to consider the **interaction** between two explanatory variables. We first described the concept of interaction between two factors in Chapter 26 for two-way ANOVA. In multiple linear regression, interaction between variables x_1 and x_2 appears as a product term x_1x_2, making a new variable x_3 in the model. The product term means that *the relationship between the mean response and one explanatory variable* x_1 *changes when we change the value of the other explanatory variable* x_2. Here is an example.

EXAMPLE 28.12 Lung capacity in boys and girls

STATE: Children’s respiratory health has become a major concern due to the effects of airborne pollution and secondhand smoke. To understand the impact of these negative factors, however, it is important first to understand the respiratory system of healthy children. A common way to assess respiratory function is by measuring the forced expiratory volume (FEV), which represents the amount of air that can be actively exhaled in one second. A study recorded the FEV of a random sample of 589 boys and girls aged 3 to 19 years to determine how FEV is related to age and gender in children.[8]

PLAN: Fit and evaluate a model with two regression lines for predicting FEV.

Since boys and girls have different growth patterns, it is likely that the pattern of FEV over time will differ somewhat for boys and girls. The scatterplot in Figure 28.8 confirms this suspicion. The pattern is linear for both boys and girls, but the two lines are not parallel. Let’s see how adding an interaction term can help us model the data. Consider the following model:

$$\mu y=\beta 0+\beta 1x1+\beta 2x2+\beta 3x1x2$$

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

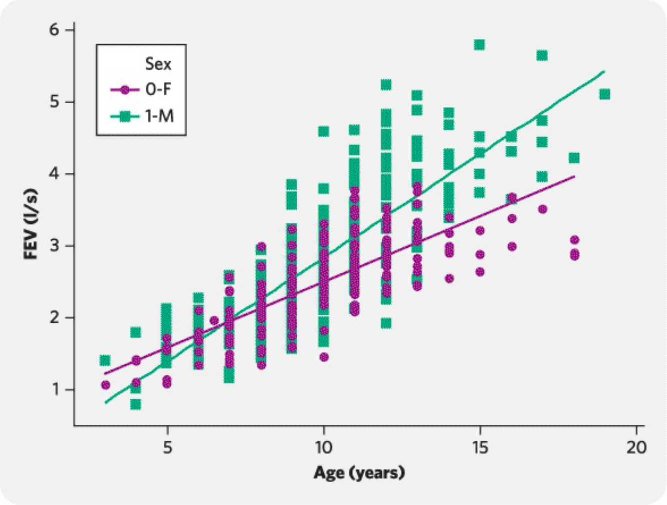


Figure 28.8
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.8 Scatterplot of FEV versus age for male and female children, for Example 28.12.

Here y is the FEV value, x_1 is the child’s age, x_2 is an indicator variable that is 1 if the child is a boy and 0 otherwise, and $x_1 x_2$ is the interaction term. For girls, $x_2 = 0$, so the model is

$$\mu y=\beta 0+\beta 1x1$$

$$\mu_y = \beta_0 + \beta_1 x_1$$

For boys, $x_2 = 1$, so the model is

$$\mu y=\beta 0+\beta 1x1+\beta 2+\beta 3x1=(\beta 0+\beta 2)+(\beta 1+\beta 3)x1$$

$$\begin{aligned}\mu_y &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1\end{aligned}$$

A careful look allows us to interpret all four parameters: β_0 and β_1 are the intercept and slope, respectively, for girls; β_2 and β_3 indicate the fixed change in the intercept and slope, respectively, for boys. Be careful not to interpret β_2 as the intercept and β_3 as the slope for boys. The indicator variable allows us to change the intercept as we did before, and the new interaction term allows us to change the slope. Thus, if β_3 is zero, the two lines are parallel. If β_3 is nonzero, then the lines are not parallel. The test for $H0:\beta3=0$ $H_0:\beta_3=0$ is sometimes called the "test of parallelism."

INTERACTION IN A MODEL WITH TWO REGRESSION LINES

We have *n* observations on an explanatory variable x_1, an indicator variable x_2 coded as 0 for some individuals and as 1 for other individuals, and a response variable *y*. The mean response μ_y is a linear function of the four parameters β_0, β_1, β_2, and β_3:

$$\mu y=\beta 0+\beta 1x1+\beta 2x2+\beta 3x1x2$$

$$\mu_y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$$

EXAMPLE 28.13 Lung capacity in boys and girls

SOLVE: Figure 28.8 shows the two regression lines, one for girls and one for boys. The fitted model, as shown by the two regression lines in this case, appears to provide a good visual summary for the two groups.

Figure 28.9 provides the regression output from Minitab. By substituting 0 and 1 for the indicator variable x_2, we can easily obtain the two estimated regression lines. The estimated regression lines are

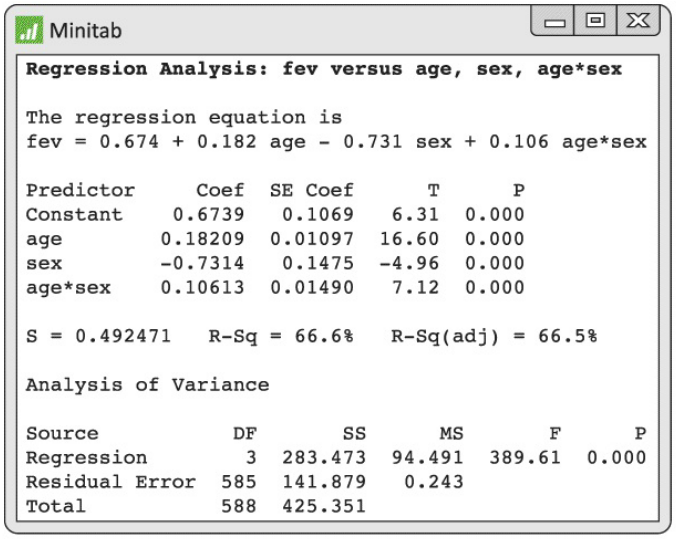
$$y\text{^}=0.674+0.182x1 \quad \text{for girls}$$

$$\hat{y}=0.674+0.182x_1 \quad \text{for girls}$$

and

$$y\text{^}=(0.674-0.731)+(0.182+0.106)x1=-0.057+0.288x1 \quad \text{for boys}$$

$$\hat{y}=(0.674-0.731)+(0.182+0.106)\,x_1=-0.057+0.288x_1 \quad \text{for boys}$$



```
Regression Analysis: fev versus age, sex, age*sex

The regression equation is
fev = 0.674 + 0.182 age - 0.731 sex + 0.106 age*sex

Predictor      Coef  SE Coef      T      P
Constant     0.6739   0.1069   6.31  0.000
age         0.18209  0.01097  16.60  0.000
sex         -0.7314   0.1475  -4.96  0.000
age*sex     0.10613  0.01490   7.12  0.000

S = 0.492471   R-Sq = 66.6%   R-Sq(adj) = 66.5%

Analysis of Variance

Source            DF       SS      MS       F      P
Regression         3  283.473  94.491  389.61  0.000
Residual Error   585  141.879   0.243
Total            588  425.351
```

Figure 28.9
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.9 Output from Minitab for the model with two regression lines in Example 28.13.

The overall *F* statistic 389.61 and the corresponding *P*-value in the ANOVA table clearly indicate that the model with the two variables and their interaction is helpful in predicting FEV. Looking at the individual *t* tests for the coefficients, we notice that all are highly significant and the model explains 66.6% of variations in FEV ($R^2 = 66.6\%$). Therefore, all three variables are useful in explaining FEV, and the two regression lines (for boys and for girls) are not parallel. That is, the lung capacity of boys and girls does not develop similarly as children grow up. Lung capacity increases faster in boys.

The residual plots (not shown) indicate no major deviation from Normality or linearity, but possibly some minor concern with the assumption of constant variance.

CONCLUDE: The model with two regression lines, one for girls and one for boys, and interaction between age and gender explains approximately 67% of the variation in FEV values. This model provides a better fit than the simple linear regression model predicting FEV from age alone ($R^2 = 61.1\%$, not shown).

Even though, for simplicity, we have discussed models without interaction first, it is better in practice to consider models with interaction *terms before going to the more restrictive model with parallel regression lines. If you begin your model fitting with the more restrictive model with parallel regression lines, you are basically assuming that there is no interaction.* We won’t discuss model selection formally, but deciding which model to use is an important skill.

CAUTION

The concept of interaction is fairly simple to visualize when there are only two variables and one of them is an indicator variable because we can simply see whether the two lines are parallel or not. But *multiple regression* models can include any number of explanatory variables, and interaction need not necessarily involve a simple indicator variable. In fact, the multiple linear regression model can even include squares or higher powers of quantitative explanatory variables, as long as the model is *linear* and each variable is multiplied by a constant β. Here are some conceptual examples that illustrate the flexibility of multiple regression models.

EXAMPLE 28.14 Two interacting explanatory variables

Suppose that we have n observations on two continuous explanatory variables x_1 and x_2 and a response variable y. Our goal is to predict the behavior of y for given values of x_1 and x_2 such that the mean response is given by

$$\mu_y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$$

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Because there are two explanatory variables x_1 and x_2, we can graph the relationship of y with x_1 and x_2 in three dimensions. Figure 28.10 shows y vertically above a plane in which x_1 and x_2 take their values. The result is a surface in space. Figure 28.10(a) shows the easiest extension of our simple linear regression model from Chapter 23. Instead of fitting a line to the data, we are now fitting a plane. This figure shows the plane $\mu_y=x_1+x_2$. The plane is a population model, and when we collect data on our explanatory variables, we will see vertical deviations from the points to the plane. The goal of our least-squares regression model is to minimize the vertical distances from the points to the plane.

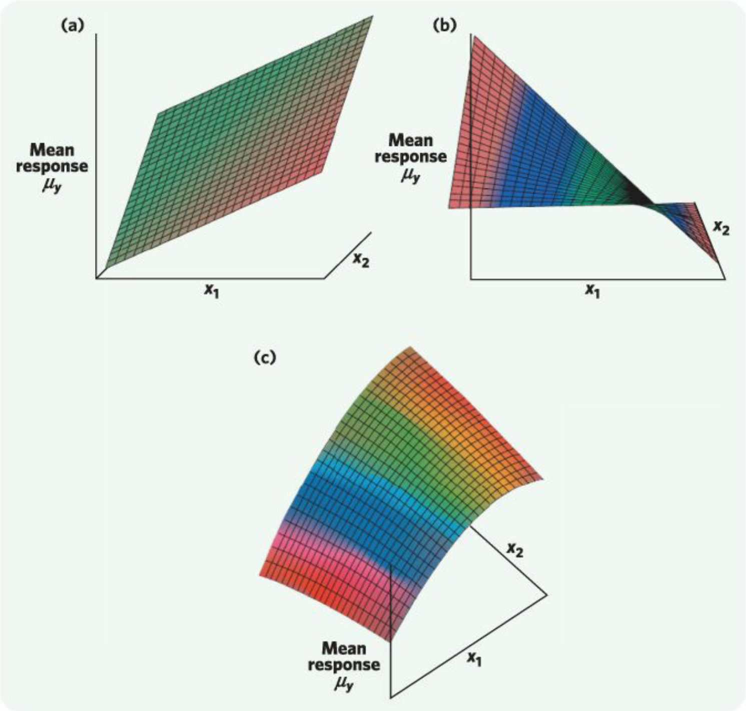


Figure 28.10
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.10 Some possible surfaces for multiple regression models, for Example 28.14.

(a) Plane μy=x1+x2. $\mu_y = x_1 + x_2$.

(b) Surface μy=2x1+2x2+10x1x2.

$\mu_y = 2x_1 + 2x_2 + 10x_1x_2$. (c) Surface μy=2000-20x12-2x1-3x22+5x2+10x1x2.

$\mu_y = 2000 - 20x_1^2 - 2x_1 - 3x_2^2 + 5x_2 + 10x_1x_2$.

Figure 28.10(b) adds a slight twist created by the interaction term in the model. The mean response in Figure 28.10(b) is μy=2x1+2x2+10x1x2.$\mu_y = 2x_1 + 2x_2 + 10x_1x_2$. The coefficients in front of the explanatory variables indicate part of the effect of a 1-unit change on the mean response for each 1-unit change in one of the explanatory variables. But the interpretation of the effect of a 1-unit change on the mean response for one variable also depends on the other variable. For example, if $x_2 = 1$, the mean response increases by 12 (μy=2x1+2(1)+10x1(1)=2+12x1$\mu_y = 2x_1 + 2(1) + 10x_1(1) = 2 + 12x_1$) for a 1-unit increase in x_1. However, when $x_2 = 2$, the mean response increases by 22 (μy=2x1+2(2)+10x1(2)=4+22x1$\mu_y = 2x_1 + 2\,(2) + 10x_1\,(2) = 4 + 22x_1$) for a 1-unit increase in x_1. *To*

interpret the parameters in multiple regression models, we think about the impact of one variable on the mean response while all the other variables are held fixed.

Another way to think about possible changes in the mean response for different possible multiple regression models for two explanatory variables is to take a piece of paper and hold it as shown in Figure 28.10(a). Now begin moving the corners of the paper to get different surfaces. You will discover that a wide variety of surfaces are possible with only two explanatory variables.

Another possible response surface is shown in Figure 28.10(c). A quick inspection of this figure reveals some curvature in the mean response. Multiple linear regression models are linear in the parameters, so we can fit quadratic models by squaring the explanatory variables, or higher-order polynomial models by considering higher-order terms for the explanatory variables. However, these polynomial models require special care, beyond the scope of this textbook. The mean response in Figure 28.10(c) is μy=2000-20x12-2x1-3x22+5x2+10x1x2. $\mu_y = 2000 - 20x_1^2 - 2x_1 - 3x_2^2 + 5x_2 + 10x_1x_2$. Notice that this mean response has two linear terms, two quadratic terms, and one interaction term. Models of this form are known as second-order polynomial regression models.

Software estimates the parameters just as before, finding the β's by the least-squares method and estimating σ by the regression standard error based on the residuals. Nothing is new there except more complicated calculations that software does for us and, of course, much more challenging interpretations.

One of the greatest practical challenges lies with regression coefficients because *the*

relationship between the response y and any one explanatory variable can change greatly, depending on what other explanatory variables are present in the model. In fact, when the explanatory variables are correlated, multiple regression models can produce some very odd and counterintuitive results, so it is important to check carefully for correlation among a potential set of explanatory variables. We shall see a concrete example of this in the following case study.

APPLY YOUR KNOWLEDGE

28.11 Physiology of men and women. High blood pressure and arterial stiffness are risk factors for cardiovascular disease, the leading cause of death worldwide. The sympathetic nervous system is known to affect vasoconstriction. Researchers investigated the relationship between sympathetic nerve activity (in number of bursts per 100 heartbeats) and an indirect measure of arterial stiffness (augmented aortic pressure, in mm Hg) in 44 healthy young adults. The results are displayed in Table 28.2.[9]

a. Plot arterial stiffness versus sympathetic nerve activity using different symbols for men and women. Does a multiple regression model with parallel slopes for men and women make sense? Explain.
b. Create a new variable representing the interaction x_1x_2 between the two explanatory variables. Then fit a model that includes both explanatory variables and their interaction. What do you conclude?
c. Are the conditions for inference met for your model in part b? Construct appropriate residual plots and comment.

TABLE 28.2 Sympathetic nerve activity (bursts/100 heartbeats) and arterial stiffness (mm Hg) of men and women

IndMen	Sympathetic	Stiffness	IndMen	Sympathetic	Stiffness
1	12.8	−2.0	0	17.5	10.0
1	14.2	−1.1	0	20.1	10.0
1	17.7	−1.0	0	18.8	8.3
1	19.9	0.0	0	18.1	7.0
1	31.1	0.5	0	16.9	4.9
1	34.6	2.0	0	14.3	4.5

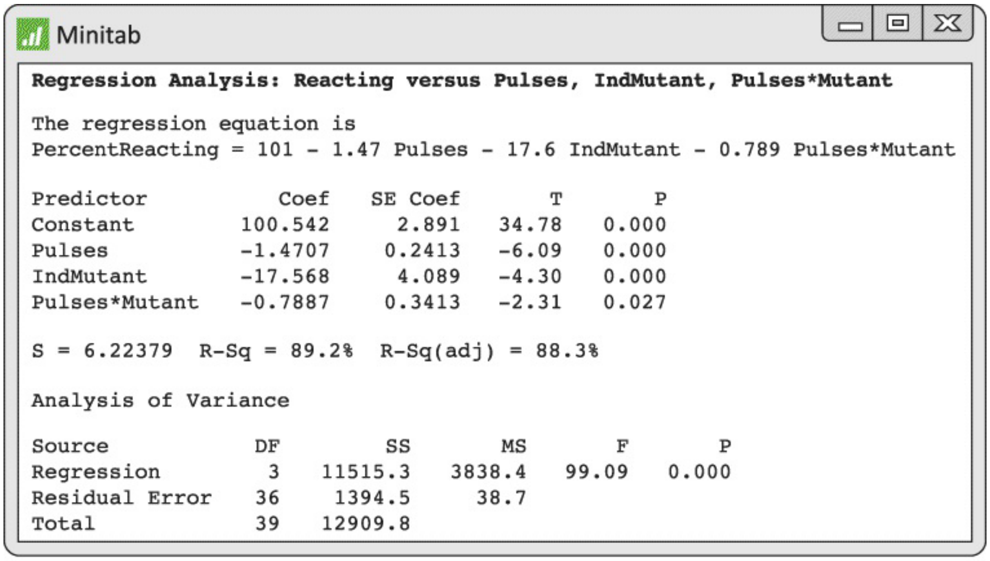
1	35.9	2.0	0	9.9	4.0
1	37.9	2.0	0	9.6	3.4
1	45.0	8.0	0	17.3	0.0
1	46.1	2.7	0	17.9	2.0
1	66.7	7.5	0	22.5	2.5
1	63.8	0.0	0	23.4	3.5
1	47.2	−4.0	0	23.4	4.3
1	36.5	0.0	0	27.7	3.0
1	34.7	−4.0	0	27.1	2.0
1	33.3	−4.5	0	34.2	−0.5
1	28.2	−5.5	0	34.9	1.5
1	29.1	−2.9	0	36.9	4.0
1	29.8	−2.5	0	37.1	4.4
1	24.1	−3.0	0	37.5	8.0
1	18.4	−5.5	0	49.2	−3.0
1	5.3	9.0	0	52.7	−4.0

28.12 Neural mechanism of nociception: interaction. You can add an interaction term to test whether the lines in a model with two lines are likely to be parallel. Exercise 28.9 described the changes in percent reacting to light pulses in two subpopulations of *C. elegans*. The scatterplot created for part (a) of that exercise showed two regression lines with negative slopes.

a. Create a new variable equal to the product of Pulses and IndMutant. This is your interaction term, Pulses*Mutant. Then use software to create a multiple linear regression model of percent reacting as a function of Pulses, IndMutant, and Pulses*Mutant. Compare it with the Minitab output provided below and make sure that your results are similar, up to rounding errors.

b. Interpret the software output for the model with interaction. Is the model significant overall? What percent of the variations in percent reacting is explained by the model?

c. Obtain the equation of the estimated regression line for each subpopulation. Are the slopes significantly different for the two subpopulations? Explain your answers.



```
Regression Analysis: Reacting versus Pulses, IndMutant, Pulses*Mutant

The regression equation is
PercentReacting = 101 - 1.47 Pulses - 17.6 IndMutant - 0.789 Pulses*Mutant

Predictor         Coef    SE Coef       T       P
Constant       100.542      2.891   34.78   0.000
Pulses         -1.4707     0.2413   -6.09   0.000
IndMutant      -17.568      4.089   -4.30   0.000
Pulses*Mutant  -0.7887     0.3413   -2.31   0.027

S = 6.22379  R-Sq = 89.2%  R-Sq(adj) = 88.3%

Analysis of Variance

Source          DF        SS       MS      F       P
Regression       3   11515.3   3838.4  99.09   0.000
Residual Error  36    1394.5     38.7
Total           39   12909.8
```

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

A CASE STUDY FOR MULTIPLE REGRESSION

We will now illustrate step-by-step the process of arriving at a suitable multiple regression model from raw data. To build a multiple regression model, first examine the data for outliers and other deviations that might unduly influence your conclusions. Next, use descriptive statistics, especially correlations, to get an idea of which explanatory variables may be most helpful in explaining the response and to find out if some of the explanatory variables are also correlated. Fit several models using combinations of these variables, paying attention to the individual *t* statistics to see if any variables contribute little in this particular model.

Always think about the real-world setting of your data and use common sense as part of the process. When building a regression model, it is far more efficient to start with a selection of variables that make sense biologically, than to search systematically through a large number of variables in a blind fashion.

EXAMPLE 28.15 Energy cost of running: exploratory analysis

Tom Tracy Photography/Alamy

If you run on a computerized treadmill at the gym and check the calories you have just burned, chances are that the machine's answer is based on the following 1963 study. Researchers asked athletes to run on a treadmill at various speeds and inclines and assessed the athletes' energy expenditure (computed indirectly via oxygen consumption and individual body measurements). Table 28.3 provides data from 25 measurements taken under various conditions of speed (in kilometers per hour) and treadmill incline (in percent).[10] We want to find a model that can explain the variations in energy expenditure.

TABLE 28.3 Energy expenditure while running at various speeds (km/h) and treadmill inclines (%)

Speed	Energy	Incline	Speed	Energy	Incline
8.7	3.0	−10	13.8	13.5	0
11.3	4.4	−10	18.9	18.1	0
13.8	6.2	−10	10.0	13.5	5
16.3	7.8	−10	8.8	12.2	5
18.8	8.4	−10	6.3	9.1	5
21.4	10.0	−10	5.0	10.6	10
6.2	3.7	−5	6.9	13.1	10
8.8	5.9	−5	7.5	14.3	10
13.8	9.1	−5	8.8	16.7	10
16.3	12.0	−5	4.4	14.0	15
6.2	6.1	0	5.0	15.0	15
8.8	8.4	0	6.0	16.8	15
11.3	11.1	0			

The response variable y is energy expenditure. A careful examination of these values reveals no outliers or skew. We plot the energy values against speed and against incline in two separate scatterplots. They are shown in Figure 28.11(a) and (b), respectively. As expected, the treadmill's incline influences energy expenditure. This relationship is positive and linear. Surprisingly though, there is no direct relationship between energy expenditure and running speed in the data set.

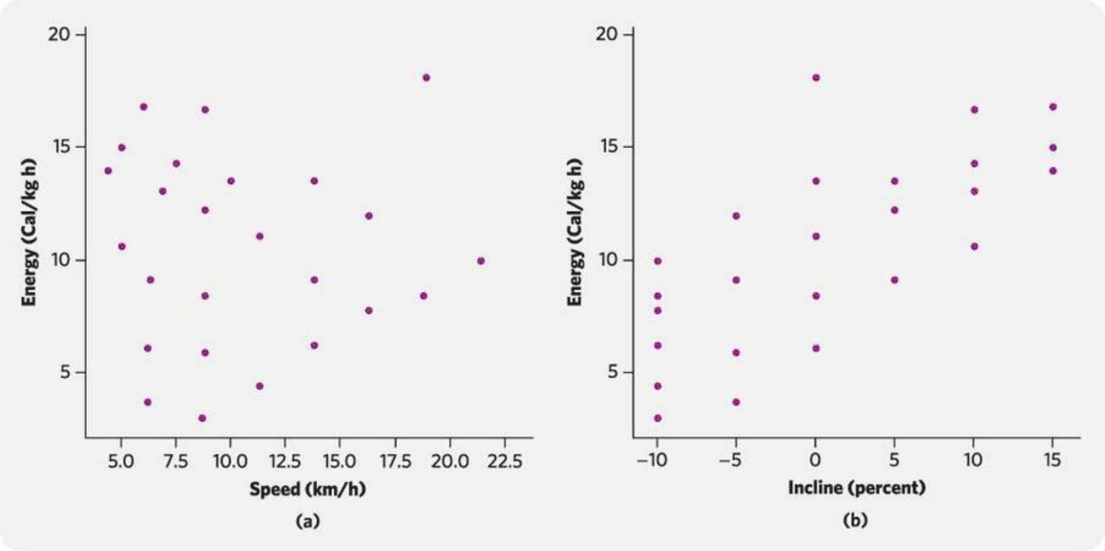


Figure 28.11
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.11 Scatterplots of energy expenditure against (a) speed and against (b) incline for Example 28.15.

A correlation analysis gives the following correlation coefficients and associated *P*-values:

Energy and Incline	$r = 0.726$, $P < 0.001$
Energy and Speed	$r = -0.022$, $P = 0.916$
Incline and Speed	$r = -0.686$, $P < 0.001$

The correlation analysis confirms what we saw in the scatterplots, but it also reveals that incline and speed are strongly correlated ($r = -0.686$). For the significant relationship between energy and incline, software gives the following simple linear regression model:

y^=10.2+0.352 Incline

$$\hat{y} = 10.2 + 0.352\ Incline$$

The exploratory analysis shows that, in this particular context when *both* speed and incline are varied, speed alone cannot explain the observed variations in energy expenditure. Incline alone, on the other hand, explains 53% (R2=0.527$R^2 = 0.527$) of the variations in energy expenditure. Could we obtain an even better model?

The finding that speed and incline are correlated is also very important. It tells us that the effect of one may be offset or modified by the other. Therefore, although the relationship between energy and speed is not significant, we shouldn't discount speed in building a model of running energy expenditure. Also, if common sense or previous studies suggest that a variable may be influential, you should make sure to study it carefully. Common sense certainly points to some impact of running speed on energy expenditure.

EXAMPLE 28.16 Energy cost of running: displaying more complex relationships

One of the challenges of multiple regression is visualizing complex relationships. At this point, we would like to see how energy expenditure relates to *both* speed and incline. Figure 28.12(a) shows a 3D scatterplot of the three variables with a surface fit. The dots on this graph follow a clear pattern that depends on both speed and incline, supporting our suspicion that speed should impact energy expenditure in some way. If your software program creates 3D plots, try rotating the graph to better see the pattern.

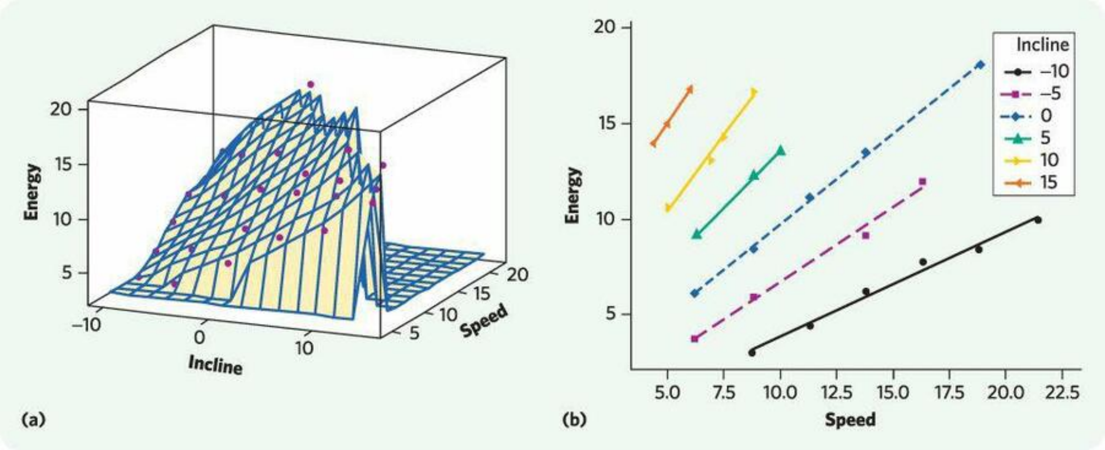


Figure 28.12
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.12 Scatterplots of energy expenditure against speed and incline using (a) a 3D-surface plot or a (b) 2D plot with regression lines of different colors for various inclines.

In our case, incline is discrete, taking only 6 possible values. Therefore, we can create a scatterplot of energy against speed and color-code the dots for each incline value. Figure 28.12(b) shows such a plot. The relationship between energy, speed, and incline is now much clearer. For any fixed incline value, energy expenditure increases linearly with speed. For similar speeds, energy expenditure is higher for higher inclines. We also notice that the lines corresponding to different inclines appear to be somewhat steeper for higher inclines than for lower ones.

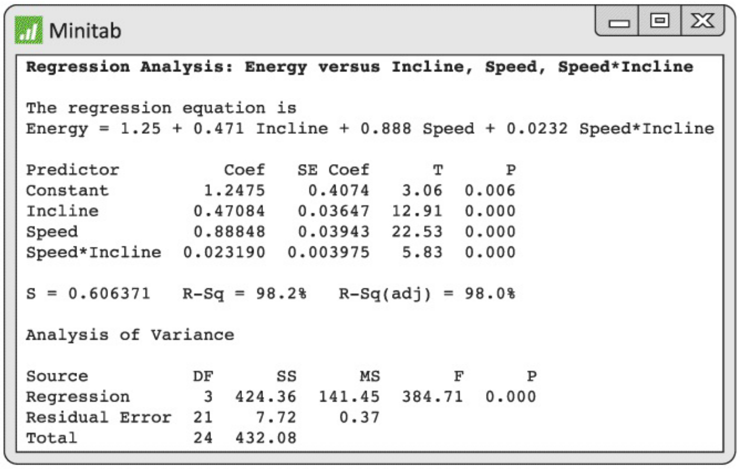
The more advanced graphing led us to expose the relationship between speed and energy in this experiment, even though we found that the correlation between the two is not significant. Therefore, we definitely want to include both speed and incline to build a model of energy expenditure. Because the lines in Figure 28.12(b) are not all parallel, we will also check whether an interaction term between speed and incline helps predict energy.

EXAMPLE 28.17 Energy cost of running: comparing models

We create an interaction variable x_1x_2, *Speed*Incline*. Figure 28.13 shows the multiple regression output using *Speed*, *Incline*, and their interaction *Speed*Incline*. The overall model is significant (*P*-value of ANOVA test < 0.0005), and residual plots (not shown) give no reason to discard the model. All regression coefficients in the model are significant too (P<0.0005$P < 0.0005$) and, together, they explain 98.2% of the variations in energy expenditure. The fitted model using all three explanatory variables is

y^=1.25+0.471 Incline+0.888 Speed+0.0232 Speed*Incline

$$\hat{y} = 1.25 + 0.471\,\mathit{Incline} + 0.888\,\mathit{Speed} + 0.0232\,\mathit{Speed}^{*}\mathit{Incline}$$



```
Regression Analysis: Energy versus Incline, Speed, Speed*Incline

The regression equation is
Energy = 1.25 + 0.471 Incline + 0.888 Speed + 0.0232 Speed*Incline

Predictor          Coef    SE Coef      T      P
Constant         1.2475     0.4074   3.06  0.006
Incline         0.47084    0.03647  12.91  0.000
Speed           0.88848    0.03943  22.53  0.000
Speed*Incline  0.023190   0.003975   5.83  0.000

S = 0.606371   R-Sq = 98.2%   R-Sq(adj) = 98.0%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       3  424.36  141.45  384.71  0.000
Residual Error  21    7.72    0.37
Total           24  432.08
```

Figure 28.13

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.13 Output from Minitab for the model with two variables and their interaction term in Example 28.17.

The model using speed, incline, and their interaction is outstanding and does a much better job of explaining energy expenditure than the model using incline alone, discussed in Example 28.15. The next question is whether a simpler model could be almost as successful. (We know that removing an explanatory variable can only reduce R^2.)

Energy versus Speed, Incline	$R^2 = 95.3\%$
Energy versus Speed, Speed*Incline	$R^2 = 84.0\%$
Energy versus Incline, Speed*Incline	$R^2 = 55.0\%$

Two of the models have a drastically reduced squared multiple correlation coefficient. However, the model with speed and incline but no interaction term comes quite close, with $R^2 = 95.3\%$. The fitted model using only speed and incline is

y^=1.77+0.782 Speed+0.651 Incline

$$\hat{y} = 1.77 + 0.782\, Speed + 0.651\, Incline$$

with significant regression coefficients for both speed and incline (P<0.0005$P < 0.0005$).

While a model with only speed and incline is quite satisfactory, the full model with incline, speed, and their interaction performs substantially better. The complete model shows that, when both speed and incline are taken into account, the interaction of these two variables

> plays a significant role in explaining variations in energy expenditure. This also reflects our observation that the color-coded lines in Figure 28.12(b) are not all parallel.

Some statistical programs provide **automated algorithms** to choose regression models. All possible regression algorithms are very useful. However, automated programs do not add interaction terms, create new variables, or notice curved patterns in the residuals. *Automated algorithms that add or remove variables one at a time often miss good models.* Try building models by considering and evaluating various possible subsets of models.

It is helpful to start with clear objectives, if you can. Choose variables that fit these objectives, even in choosing the response variable. Notice that in Example 28.17, energy expenditure is measured in Calories per kilogram of body weight and per hour. The researchers could have used a simpler measure of Calories burned. However, the duration of the runs and the physiological characteristics of the runners would have added a lot of variability and made the model selection a lot more challenging. The model we have created can be used quite successfully to predict the number of Calories burned per unit of time for a person with a given weight.

APPLY YOUR KNOWLEDGE

28.13 Fish sizes. Table 28.4 contains data on the size of perch caught in a lake in Finland.[11] Use statistical software to help you analyze these data.

a. Use the multiple regression model with two explanatory variables, length and width, to predict the weight of a perch. Provide the estimated multiple regression equation.
b. How much of the variation in the weight of perch is explained by the model in part a?
c. Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of

perch? Explain.

d. Do the individual *t* tests indicate that both β_1 and β_2 are significantly different from zero? Explain.

e. Create a new variable, called interaction, that is the product of length and width. Use the multiple regression model with three explanatory variables—length, width, and interaction—to predict the weight of a perch. Provide the estimated multiple regression equation.

f. How much of the variation in the weight of perch is explained by the model in part e?

g. Does the ANOVA table indicate that at least one of the explanatory variables is helpful in predicting the weight of perch? Explain.

h. Describe how the individual *t* statistics changed when the interaction term was added.

Design Pics Inc./Alamy

TABLE 28.4 Measurements on 56 perch

Observation number	Weight (g)	Length (cm)	Width (cm)	Observation number	Weight (g)	Length (cm)	Widt (cm
104	5.9	8.8	1.4	132	197.0	27.0	4.2
105	32.0	14.7	2.0	133	218.0	28.0	4.1
106	40.0	16.0	2.4	134	300.0	28.7	5.1
107	51.5	17.2	2.6	135	260.0	28.9	4.3
108	70.0	18.5	2.9	136	265.0	28.9	4.3
109	100.0	19.2	3.3	137	250.0	28.9	4.6
110	78.0	19.4	3.1	138	250.0	29.4	4.2
111	80.0	20.2	3.1	139	300.0	30.1	4.6
112	85.0	20.8	3.0	140	320.0	31.6	4.8

113	85.0	21.0	2.8	141	514.0	34.0	6.
114	110.0	22.5	3.6	142	556.0	36.5	6.
115	115.0	22.5	3.3	143	840.0	37.3	7.
116	125.0	22.5	3.7	144	685.0	39.0	6.
117	130.0	22.8	3.5	145	700.0	38.3	6.
118	120.0	23.5	3.4	146	700.0	39.4	6.
119	120.0	23.5	3.5	147	690.0	39.3	6.
120	130.0	23.5	3.5	148	900.0	41.4	7.
121	135.0	23.5	3.5	149	650.0	41.4	6.
122	110.0	23.5	4.0	150	820.0	41.3	7.
123	130.0	24.0	3.6	151	850.0	42.3	7.1
124	150.0	24.0	3.6	152	900.0	42.5	7.
125	145.0	24.2	3.6	153	1015.0	42.4	7.
126	150.0	24.5	3.6	154	820.0	42.5	6.
127	170.0	25.0	3.7	155	1100.0	44.6	6.
128	225.0	25.5	3.7	156	1000.0	45.2	7.
129	145.0	25.5	3.8	157	1100.0	45.5	7.
130	188.0	26.2	4.2	158	1000.0	46.0	8.1
131	180.0	26.5	3.7	159	1000.0	46.6	7.

28.14 Tasting cheddar cheese. What determines the taste of cheddar cheese? Experimenters assessed the concentration of lactic acid, acetic acid, and hydrogen sulfide (H_2S) in 30 randomly chosen pieces of cheddar cheese that were also rated for taste by a panel of food tasters (the higher the score, the tastier). Data from the study are shown in Table 28.5. “Acetic” and “H_2S” are actually log-transformed concentrations.[12]

a. For each of the three explanatory variables in turn, make a scatterplot with taste on the *y* axis and find the correlation coefficient. Which relationships are linear? Which have the strongest correlation with taste?

b. Use software to obtain the regression equation and run inference for a regression model that includes all three explanatory variables. Interpret the software output, including the meaning of the value taken by R^2.

c. The model in has one nonsignificant regression coefficient. Which explanatory variable does it describe? Now create a new regression model that excludes this explanatory variable. Interpret the software output and compare it with your findings in b.

d. Both regression coefficients in c are significant. Which explanatory variable of the two has the less significant (larger) *P*-value? Create a new regression model that excludes this explanatory variable and keeps only the more significant one. How does this last model compare with the model in c?

e. Which model best explains cheddar cheese taste? Check the conditions for inference for this model and conclude. The approach you have used here to select a multiple linear regression model is called **backward selection** because you start with all reasonable explanatory variables and gradually prune the model back until you get a satisfactory one.

TABLE 28.5 Chemical composition and taste score of aged cheddar cheese

Taste	Acetic	H_2S	Lactic	Taste	Acetic	H_2S	Lactic
12.3	4.543	3.135	0.86	40.9	6.365	9.588	1.74
20.9	5.159	5.043	1.53	15.9	4.787	3.912	1.16
39.0	5.366	5.438	1.57	6.4	5.412	4.700	1.49
47.9	5.759	7.496	1.81	18.0	5.247	6.174	1.63
5.6	4.663	3.807	0.99	38.9	5.438	9.064	1.99
25.9	5.697	7.601	1.09	14.0	4.564	4.949	1.15
37.3	5.892	8.726	1.29	15.2	5.298	5.220	1.33
21.9	6.078	7.966	1.78	32.0	5.455	9.242	1.44
18.1	4.898	3.850	1.29	56.7	5.855	10.199	2.01
21.0	5.242	4.174	1.58	16.8	5.366	3.664	1.31
34.9	5.740	6.142	1.68	11.6	6.043	3.219	1.46
57.2	6.446	7.908	1.90	26.5	6.458	6.962	1.72
0.7	4.477	2.996	1.06	0.7	5.328	3.912	1.25
25.9	5.236	4.942	1.30	13.4	5.802	6.685	1.08

54.9	6.151	6.752	1.52	5.5	6.176	4.787	1.25

LOGISTIC REGRESSION

So far, we have examined how a quantitative response variable can be expressed as a linear function of any number of either quantitative or categorical explanatory variables. A simple linear regression model, for instance, represents the mean response μ_y for a given value of the explanatory variable *x* in the form of $\mu y=\beta 0+\beta 1x.$$\mu_y = \beta_0 + \beta_1 x.$ Can we apply the same principles to a categorical response variable?

We have seen in Chapters 12 and 19 that when a variable *y* is categorical, we can describe it in a binary fashion such that any outcome is either a success or a failure (arbitrarily defined). For example, we can ask whether a patient's health improves after treatment (success) or not (failure). For any given value of an explanatory variable *x*, the mean response is the population proportion of successes $py=p$$p_y = p$ for that particular value of *x*. This is also the probability of randomly obtaining a success when *x* takes this particular value.

If there is no relationship between *x* and *y*, the mean response is simply *p* for any value of *x*. But what if there is an association between *x* and *y*? Can we find a simple linear model that expresses *p* as a function of *x*? Unfortunately, this is rather unlikely. Indeed, as a probability or a proportion, *p* can take only values between 0 and 1. On the other hand, a linear function of a quantitative variable *x* is likely to create values over a whole range of intervals, especially for extreme values of *x*. So the relationship between *p* and a quantitative variable *x* is unlikely to be truly linear.

A number of examples in this and other chapters on regression discussed relationships that are not linear. It is sometimes possible to reveal a linear relationship by using a simple mathematical transformation, such as computing the logarithm of one or both quantitative variables. We now apply the same idea to describing relationships in which the response variable is categorical.

In Chapter 9 we discussed the concept of odds in probability. Odds are the ratio of two probabilities where the numerator is the probability *p* of an event and the denominator is the complementary probability of that event not occurring. That is, odds=p1-p

$$\text{odds} = \frac{p}{1-p}$$

Provided that neither probability is equal to zero, odds can take any strictly positive value. In turn, the natural logarithm of odds can take any real value, positive or negative. We can therefore hope to find a linear relationship between the natural logarithm of odds (called log odds, logit, or *L*) and some relevant continuous random variable *x*.

LOGISTIC REGRESSION

The **logistic regression model** is

L=lnp1-p=β0+β1x

$$L = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where *p* is the proportion or probability of a given outcome in a population, *x* is an explanatory variable, and *L* is the natural logarithm of the odds of that outcome in the population.

The model can be extended to include *k* explanatory variables, such that

L=lnp1-p=β0+β1x+···+βkxk

$$L = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \cdots + \beta_k x_k$$

As with the relationship between two quantitative variables, not all (*L*, *x*) pairs of variables studied will show a linear trend, but those that do can be modeled and studied with the tools of linear regression. Note also that the logistic model does not represent values of *p* that are either 0 or 1. Instead, it provides models in which *p* can come arbitrarily close to 0 or 1.

Figure 28.14(a) illustrates how a linear relationship between *L* and *x* translates when *p* is expressed as a function of *x*. Figure 28.14(b) shows that the logistic regression model corresponds to a relationship between *p* and *x* that has an “S,” or “sigmoidal,” shape. In practice, you may observe only a subset of that sigmoidal curve if your explanatory variable *x* takes values over a limited interval only. We can derive the equation for that sigmoidal curve by using the antilog transformation:

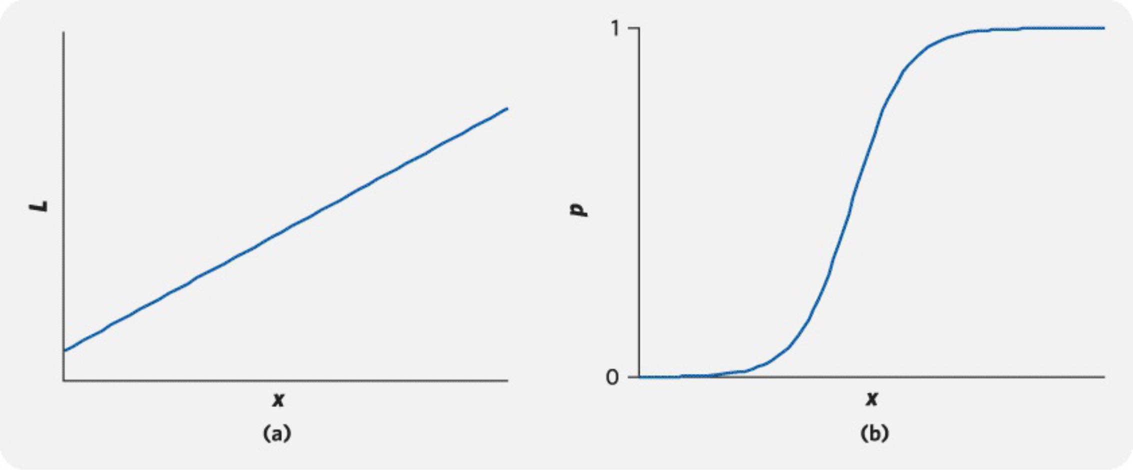


Figure 28.14

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.14 (a) Linear relationship between the logit *L* and an explanatory variable *x* and (b) its equivalent when *L* is untransformed to obtain *p*, the probability of a given outcome. The relationship between *p* and *x* has a sigmoidal, or "S," shape.

p1-p=eln(p⁄(1-p))=eβ0+β1xp=eβ0+β1x1+eβ0+β1x

$$\frac{p}{1-p} = e^{\ln(p/(1-p))} = e^{\beta_0+\beta_1 x}$$

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Therefore, there are three versions of the same model:

(1)L=ln(odds)=β0+β1x(2)odds=p1-p=eβ0+β1x(3)p=eβ0+β1x1+eβ0+β1x

$$(1) \quad L = \ln(\text{odds}) = \beta_0 + \beta_1 x$$

$$(2) \quad \text{odds} = \frac{p}{1-p} = e^{\beta_0+\beta_1 x}$$

$$(3) \quad p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

We are often more interested in the odds (2) or the probability *p* of success (3), but we first estimate the logit model (1) because of its linear property. The mathematical process needed to obtain parameter estimates for this linear model is somewhat computationally intensive. Explaining these computations is beyond the scope of an introductory statistics textbook, and we will rely solely on technology to obtain these estimates.

EXAMPLE 28.18 Diagnosing meningitis

Meningitis is a generic term that describes an inflammation of the meninges, the outer membranes protecting the brain. The source of the inflammation is usually either a viral or a bacterial infection. Meningitis is potentially deadly and must be treated promptly.

Researchers examined the test results of 352 patients with acute meningitis that were later unambiguously identified as either a viral or a bacterial infection.[13] They defined the variable *abm* (acute bacterial meningitis) as a binary response variable (y) that equals 0 when the infection is viral and equals 1 when the infection is bacterial. To help with diagnosis, the immune system's response to infection can be assessed by examining the white cell count per cubic millimeter of cerebrospinal fluid filling the meninges. Does this white cell count, *wcsf (x)*, help predict whether a case of acute meningitis is viral or bacterial?

Figure 28.15 shows dotplots of *wcsf* for viral (*abm* = 0) and for bacterial (*abm* = 1) meningitis cases. The two plots are stacked and share a common x axis, making the comparison particularly easy. The two distributions overlap, but all individuals who had a high white cell count (wcsf>500$wcsf > 500$) had bacterial meningitis. On the other hand, all patients with viral meningitis had relatively low white cell counts in the cerebrospinal fluid. Therefore, there is clearly some relationship between *abm* and *wcsf* .

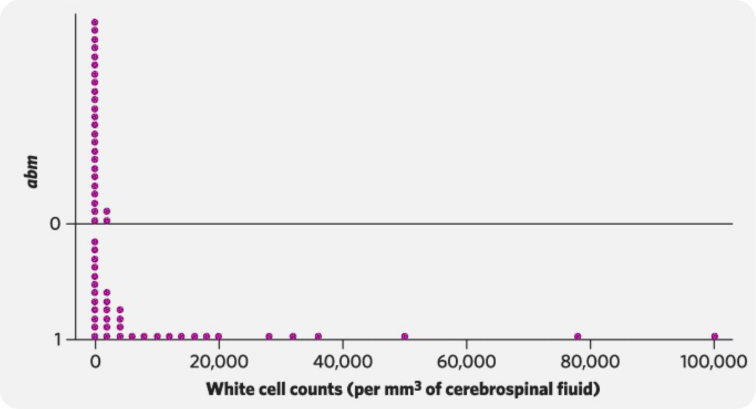
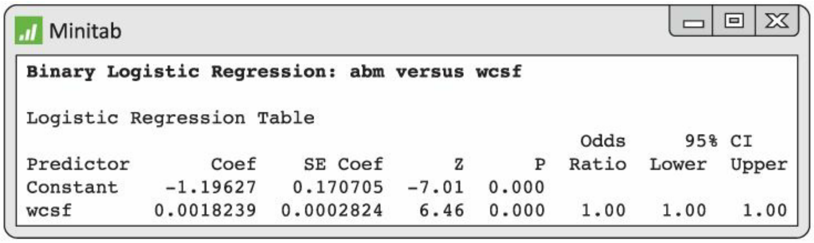


Figure 28.15
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.15 Dotplots of white cell counts (per mm^3 of cerebrospinal fluid) in patients with viral (abm = 0) or bacterial (abm = 1) meningitis, for Example 28.18. Note that each dot represents up to seven observations.

Figure 28.16 shows part of Minitab's output for the logistic regression of *abm* as a function of *wcsf* . The estimates for β_0 and β_1 are −1.1963 and 0.0018, respectively, so that the fitted logistic regression model is



```
Minitab
Binary Logistic Regression: abm versus wcsf

Logistic Regression Table
                                                  Odds     95% CI
Predictor       Coef    SE Coef      Z      P   Ratio  Lower  Upper
Constant    -1.19627   0.170705  -7.01  0.000
wcsf       0.0018239  0.0002824   6.46  0.000   1.00   1.00   1.00
```

Figure 28.16
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.16 Minitab output for the simple logistic regression model in Example 28.18.

L=lnp1-p=-1.1963+0.0018x

$$L = \ln\left(\frac{p}{1-p}\right) = -1.1963 + 0.0018x$$

If we convert this equation to express the estimated probability *p* that a given case of acute meningitis is due to a bacterial infection as a function of the white cell count per cubic millimeter *x*, we find that

p=e-1.1963+0.0018x1+e-1.1963+0.0018x

$$p = \frac{e^{-1.1963+0.0018x}}{1+e^{-1.1963+0.0018x}}$$

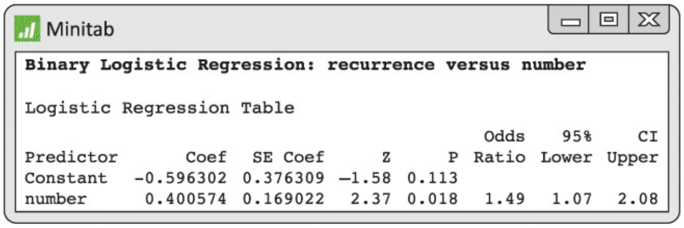
The logistic regression model allows us to estimate the probability *p* that a given case is due to bacterial infection for different test result values of *x*. Here are some results:

***x*(*wcsf*)**	0	100	500	1000	10,000
***p*(bacterial)**	0.2321	0.2657	0.4265	0.6465	0.99999995

Low white cell counts in the cerebrospinal fluid indicate that a viral infection is more likely, whereas high counts indicate that a bacterial infection is more likely, and very high counts indicate that a bacterial infection is almost certain.

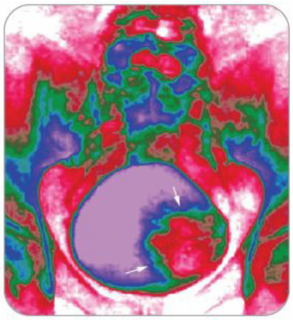
APPLY YOUR KNOWLEDGE

28.15 Recurring bladder tumors. Why do some cancers come back after a period of remission while others don't? Can we predict which cancers are more likely to recur? Researchers examined 86 patients with bladder cancer, looking for a relationship between cancer recurrence y (0 no, 1 yes) and the number x of superficial bladder tumors identified and removed right after the initial cancer diagnosis.[14] Use the Minitab output for the logistic regression analysis to find the linear equation linking the estimated logit L to the response variable x.



```
Binary Logistic Regression: recurrence versus number

Logistic Regression Table
                                                    Odds   95%    CI
Predictor       Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant   -0.596302  0.376309  -1.58  0.113
number      0.400574  0.169022   2.37  0.018   1.49   1.07   2.08
```

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



James Cavallini/Science Source

28.16 Recurring bladder tumors, continued. Return to Exercise 28.15. We are interested in the probability p that a patient with x number of initial tumors will experience a cancer recurrence. Convert the logistic equation so that you express p as a function of x. Then use this equation to estimate the probability p of cancer recurrence when $x = 0$, when $x = 2$, and when $x = 6$.

INFERENCE FOR LOGISTIC REGRESSION

Inference for logistic regression is very similar to inference for regression with a continuous response variable, except that the actual computations are more complex. In particular, the sample statistic follows approximately a Normal distribution only when the sample size is large enough.[15] (This is again an issue related to the central limit theorem we first discussed in Chapter 13.) We will rely on software for the computations and focus on interpreting the results.

INFERENCE FOR THE SLOPE OF THE LOGISTIC REGRESSION MODEL

To **test the hypothesis** $H0:β1=0,$$H_0 : \beta_1 = 0,$ compute the test statistic

$$z=\text{parameter estimate}\text{standard error of estimate}=b1SEb1$$

$$z = \frac{\text{parameter estimate}}{\text{standard error of estimate}} = \frac{b_1}{\text{SE}_{b1}}$$

Approximate *P*-values for thistest come from the standard Normal distribution.

An approximate **level C confidence interval for the slope** β_1 has the form

$$b1±z*SEb1$$

$$b_1 \pm z^* \text{SE}_{b1}$$

where z^* is the critical value for the standard Normal curve with area *C* between $-z^*$ and z^*.

EXAMPLE 28.19 Diagnosing meningitis

The output in Figure 28.16 provides parameter estimates and standard errors for the slope β_1: $b1=0.0018$$b_1 = 0.0018$ and $SEb1=0.00028.$$\text{SE}_{b1} = 0.00028.$ The test of $H0:β1=0$$H_0 : \beta_1 = 0$ indicates that the linear relationship between *L* and *x* is significant (“*P* 0.000” impliesthat

> P<0.0005$P < 0.0005$). Minitab does not provide a confidence interval for β_1, but we cancompute one very easily. A 95% confidence interval for the slope β_1 is b1±z*SEb1=0.0018±(1.96)(0.00028), $b_1 \pm z^* \mathrm{SE}_{b1} = 0.0018 \pm (1.96)\,(0.00028)$, or 0.0013 to 0.0023. That is, for every increase of 1 white cell per mm^3, the logit L increases by 0.0013 to 0.0023, with 95% confidence.

The log of odds is not an intuitive concept, and we often prefer to translate it back into odds. In particular, the ratio of two odds, or **odds ratio** (***OR***), is a commonly used comparative measure of health risk in medical science. We first discussed this concept in Chapter 20.

Suppose that the explanatory variable x increases by 1. In simple linear regression, the corresponding change in y is the slope. In logistic regression, we measure the impact of this change in the odds of success. This is the odds ratio of (x+1$x+1$) to x:

OR=oddsx+1oddsx

$$OR = \frac{\mathrm{odds}_{x+1}}{\mathrm{odds}_x}$$

Because ea⁄eb=ea-b$e^a/e^b = e^{a-b}$ and eln(a)=a,$e^{\ln(a)} = a$, it follows that

OR=oddsx+1oddsx=eβ0+β1(x+1)eβ0+β1x=eβ1

$$OR = \frac{\mathrm{odds}_{x+1}}{\mathrm{odds}_x} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}$$

An odds ratio equal to 1 indicates no effect (same risk for x and x+1$x+1$) and is equivalent to β1=0.$\beta_1 = 0$.

CONFIDENCE INTERVALS FOR THE ODDS RATIO

A **level C confidence interval for the odds ratio** eβ1e^{β_1} can be obtained by transforming the confidence interval for the slope, resulting in the interval

eb1-z*SEb1,eb1+z*SEb1

$$\left(e^{b_1 - z^* \mathrm{SE}_{b1}},\, e^{b_1 + z^* \mathrm{SE}_{b1}}\right)$$

where z^* is the critical value for the standard Normal curve with area C between $-z^*$ and z^*.

This particular odds ratio may or may not have a practical meaning. In the meningitis example, for instance, an increase of 1 white cell per mm^3 is not medically relevant and brings us no further understanding. However, when the explanatory variable in a logistic regression model is categorical, the odds ratio is commonly used to compare the odds between two conditions.

EXAMPLE 28.20 Preventing blood clots in immobilized patients

Patients immobilized for a substantial amount of time can develop deep vein thrombosis (DVT), a blood clot in a leg or pelvis vein. DVT can have serious adverse health effects and can be difficult to diagnose. On its website, drug manufacturer Pfizer reports the outcome of a study looking at the effectiveness of the drug Fragmin (dalteparin) compared with that of a placebo in preventing DVT in immobilized patients.

In a double-blind study, severely immobilized patients were randomly assigned to receive daily subcutaneous injections of either Fragmin or a placebo for two weeks. The number of patients experiencing a complication from DVT (including death) over the next 90 days is shown in the following table:

	Treatment Outcome		
	Complication	**No complication**	**Sample size**
Fragmin	42	1476	1518
Placebo	73	1400	1473

The proportions of subjects experiencing DVT complications in the two samples are

p^Fragmin=42⁄1518=0.0277p^placebo=73⁄1473=0.0496

$$\begin{aligned}\hat{p}_{\text{Fragmin}} &= 42/1518 = 0.0277\\ \hat{p}_{\text{placebo}} &= 73/1473 = 0.0496\end{aligned}$$

The odds of a subject experiencing DVT complications in the two samples are

oddsFragmin=42⁄1476≈0.02846 and ln (oddsFragmin)≈-3.5594oddsplacebo=73⁄1400≈0.05214 and ln (oddsplacebo)≈-2.9538

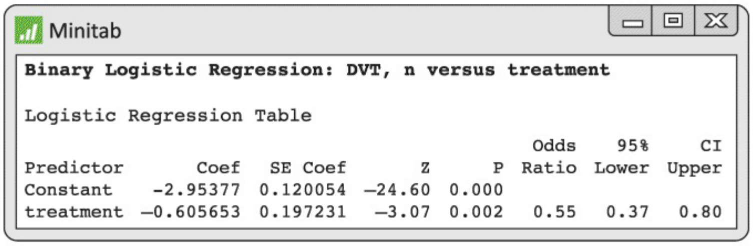
$$\begin{aligned}\text{odds}_{\text{Fragmin}} &= 42/1476 \approx 0.02846 \text{ and } \ln\left(\text{odds}_{\text{Fragmin}}\right) \approx -3.5594 \\ \text{odds}_{\text{placebo}} &= 73/1400 \approx 0.05214 \text{ and } \ln\left(\text{odds}_{\text{placebo}}\right) \approx -2.9538\end{aligned}$$

If we set the indicator variable *treatment*, or x,x, equal to 0 for the placebo group and 1 for the Fragmin group, we can derive from the above information the equation for the logistic regression model:

L=ln (oddsDVT)=ln (oddsplacebo)+[ln (oddsFragmin)-ln (oddsplacebo)]xL=ln (oddsDVT)=-2.9538+(-3.5594+2.9538)x=-2.9538-0.6056x

$$\begin{aligned}L &= \ln\left(\text{odds}_{\text{DVT}}\right) = \ln\left(\text{odds}_{\text{placebo}}\right) + \left[\ln\left(\text{odds}_{\text{Fragmin}}\right) - \ln\left(\text{odds}_{\text{placebo}}\right)\right] x \\ L &= \ln\left(\text{odds}_{\text{DVT}}\right) = -2.9538 + (-3.5594 + 2.9538)\, x = -2.9538 - 0.6056x\end{aligned}$$

Minitab gives the same result, up to rounding error, as can be seen from the output shown here:



```
Binary Logistic Regression: DVT, n versus treatment

Logistic Regression Table
                                                   Odds    95%     CI
Predictor      Coef    SE Coef       Z       P    Ratio  Lower  Upper
Constant   -2.95377   0.120054  -24.60   0.000
treatment  -0.605653  0.197231   -3.07   0.002    0.55   0.37   0.80
```

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

The estimated odds ratio of DVT complication is

oddsx=1oddsx=0=eb1=e-0.6056≈0.546

$$\frac{\text{odds}_{x=1}}{\text{odds}_{x=0}} = e^{b_1} = e^{-0.6056} \approx 0.546$$

This estimate is also the observed ratio of the two experimental odds, oddsFragmin⁄oddsplacebo≈0.02846⁄0.05214≈0.546. $\text{odds}_{\text{Fragmin}}/\text{odds}_{\text{placebo}} \approx 0.02846/0.05214 \approx 0.546.$ A 95% confidence interval for the odds ratio can be calculated to be

(eb1-z*SEb1,eb1+z*SEb1)=(e-0.6056-(1.96)(0.1972),e-0.6056+(1.96)(0.1972))=(e-0.9921,e-0.2291)≈(0.371,0.803)

$$\begin{aligned}\left(e^{b_1 - z^* \text{SE}_{b1}}, e^{b_1 + z^* \text{SE}_{b1}}\right) &= \left(e^{-0.6056-(1.96)(0.1972)}, e^{-0.6056+(1.96)(0.1972)}\right) \\ &= \left(e^{-0.9921}, e^{-0.2291}\right) \approx (0.371, 0.803)\end{aligned}$$

Our calculations agree with the Minitab output, up to rounding error. Notice that the confidence interval for *OR* is not symmetric around the estimate. This is because the exponential function is nonlinear.

The *OR* confidence interval means that, with 95% confidence, the odds of DVT complication in the Fragmin group are somewhere between 0.37 and 0.80 times the corresponding odds for the placebo group. Because we are 95% confident that *OR* is less than 1, we can conclude that the Fragmin treatment helps reduce the odds of DVT complication in immobilized patients.

APPLY YOUR KNOWLEDGE

28.17 Recurring bladder tumors, continued. Go back to Exercise 28.15 and use Minitab's output in your answers. Is the slope of the logistic regression model significant? Give the 95% confidence interval for the slope of the logistic regression. Can you conclude with 95% confidence that the slope is positive? What does this mean in terms of the impact of the number of initial tumors on the probability of cancer recurrence?

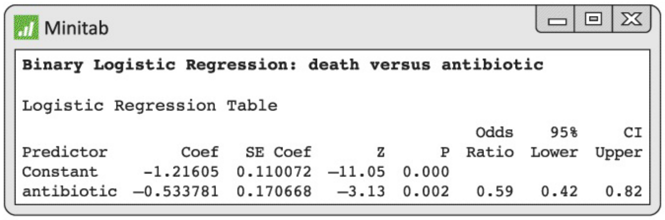
28.18 Antibiotics in the ICU. Patients who are admitted to the intensive care unit (ICU) and require mechanical ventilation can develop fatal bacterial infections. A clinical trial randomly assigned to two groups 934 patients who were admitted to an ICU and required mechanical ventilation. The control group received a standard medical treatment, whereas the antibiotics group received antibiotics prophylactically in addition to standard treatment. The study found that 107 of the 468 patients in the control group died, compared with 69 of the 466 patients in the antibiotics group.[16]

a. What are the odds of dying in each of the two groups? Use this information to compute the equation of the logistic regression model if the indicator variable *x* equals 0 for the control group and 1 for the antibiotics group.

b. Below is Minitab's logistic regression output for these data.

Show how you would compute the odds ratio for death and its 95% confidence interval. Check that your calculations agree with Minitab's.

c. Interpret your results in the context of the study.



```
Binary Logistic Regression: death versus antibiotic

Logistic Regression Table
                                                   Odds    95%     CI
Predictor        Coef   SE Coef       Z       P   Ratio  Lower  Upper
Constant     -1.21605  0.110072  –11.05   0.000
antibiotic  –0.533781  0.170668   –3.13   0.002   0.59   0.42   0.82
```

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

CHAPTER 28 SUMMARY

- An indicator variable x_2 can be used to fit a regression model with **two parallel lines.** The mean response is

  $$\mu y=\beta 0+\beta 1x1+\beta 2x2$$

  $$\mu_y=\beta_0+\beta_1x_1+\beta_2x_2$$

  where x_1 is a continuous explanatory variable.

- A multiple regression model with **two regression lines** includes a continuous variable x_1, an indicator variable x_2, and an interaction term x1x2.x_1x_2. The mean response is

  $$\mu y=\beta 0+\beta 1x1+\beta 2x2+\beta 3x1x2$$

  $$\mu_y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2$$

- The mean response μ_y for a general **multiple regression model** based on k explanatory variables x1,x2,…,xk$x_1,x_2,\ldots,x_k$ is

  $$\mu y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta kxk$$

  $$\mu_y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$$

- The **estimated regression model** is

  $$y\text{^}=b0+b1x1+b2x2+\cdots+bkxk$$

  $$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_kx_k$$

  where the b’s are obtained by the method of least squares. Use software to obtain these estimates.

- The **regression standard error** s has n-k-1$n-k-1$ degrees of freedom and is used to estimate σ.

- The **analysis of variance (ANOVA) table** breaks the total variability in the responses into two pieces. One piece summarizes the variability due to the model, and the other piece summarizes the variability due to error.

  $$\text{total variation=variation explained by model+residual (“error”) variation}$$

  $$\text{total variation} = \text{variation explained by model} + \text{residual (“error”) variation}$$

- The **squared multiple correlation coefficient** R^2 represents the proportion of variability in the response variable *y* that is explained by the explanatory variables x_1, x2,…,xk$x_2, \dots, x_k$ in a multiple regression model.

- To test the hypothesis that all the regression coefficients (β's) except β_0 are equal to zero, use the **ANOVA *F* statistic.** The null hypothesis says that the *x*'s do not help predict *y*. The alternative is that at least one of the explanatory variables is helpful in predicting *y*.
- **Individual *t* procedures** in regression inference have n-k-1$n-k-1$ degrees of freedom. These individual *t* procedures are dependent on the other explanatory variables specified in a multiple regression model. Individual *t* tests assess the contribution of one explanatory variable in the presence of the other variables in a model. The null hypothesis is written as H0:β=0$H_0 : \beta = 0$ but interpreted as "the coefficient of *x* is 0 *in this particular model*."
- **Confidence intervals** for the mean response μ_y have the form y^±t*SEμ^.$\hat{y} \pm t^* \mathrm{SE}_{\hat{\mu}}$.
- **Prediction intervals** for an individual future response *y* have the form y^±t*SEy^.$\hat{y} \pm t^* \mathrm{SE}_{\hat{y}}$.
- When the **response variable is categorical,** it can be coded as $y = 1$ for a given outcome (often arbitrarily defined as "success") and $y = 0$ otherwise ("failure"). The population mean response is the probability *p* of the given outcome (success) in the population. We use **logistic regression** methods to describe the relationship between *p* and any set of explanatory variables.
- The simple **logistic regression model** with one explanatory variable *x* is

L=ln(p1-p)=β0+β1x

$$L = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

where *L* is the natural logarithm of the odds of a given outcome in the population. The estimated model is Lestimate=b0+b1x,$L_{\text{estimate}} = b_0 + b_1 x$, which can be rewritten as

pestimate=eb0+b1x1+eb0+b1x

$$p_{\text{estimate}} = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

- A Normal *z* procedure is used to **test the hypothesis** that the slope β_1 is equal to zero and to compute an approximate **level C confidence interval for the slope** of the form b1±z*SEb1.$b_1 \pm z^* \text{SE}_{b1}$.
- Logistic regression models often report an approximate **level *C* confidence interval for the odds ratio** OR=eβ1$\boldsymbol{OR} = e^{\beta_1}$**,** which is obtained by transforming the confidence interval for the slope, giving the interval

(eb1-z*SEb1,eb1+z*SEb1)

$$\left(e^{b_1 - z^* \text{SE}_{b1}}, e^{b_1 + z^* \text{SE}_{b1}}\right)$$

STATISTICS IN SUMMARY

Here are the most important skills you should have acquired from reading this chapter.

A. Preliminaries

1. Examine the data for outliers and other deviations that might influence your conclusions.
2. Use descriptive statistics, especially correlations, to get an idea of which explanatory variables may be most helpful in explaining the response.
3. Make scatterplots to examine the relationships between explanatory variables and a response variable.
4. Use software to compute a correlation matrix to explore the relationships between pairs of variables.

B. Recognition

1. Recognize when a multiple regression model with parallel regression lines is appropriate.
2. Recognize when an interaction term needs to be added to fit a multiple regression model with two separate regression lines.
3. Recognize when a multiple regression model with several explanatory variables is appropriate.
4. Recognize the difference between the overall F test and the individual t tests.
5. Recognize when the explanatory variable is categorical, requiring the use of a logistic regression model.
6. Recognize that the parameter estimates, t statistics, and P-values for each explanatory variable depend on the specific model.
7. Inspect the data to recognize situations in which inference isn't safe: influential observations, strongly skewed residuals in a small sample, or nonconstant variation of the data points about the regression model.

C. Doing Inference Using Computer Output

1. Use software to find the estimated multiple regression model or logistic model.
2. Explain the meaning of the regression parameters (β's) in any specific multiple regression model.
3. Understand the software output for regression. Find the regression standard error, the squared multiple correlation coefficient R^2, and the overall F test and P-value. Identify the parameter estimates, standard errors, individual t tests, and P-values.
4. Use that information to carry out tests and calculate confidence intervals for the β's.
5. Use R^2 and residual plots to assess the fit of a model.
6. Choose a model by comparing R^2-values, regression standard errors, and individual t statistics.
7. Explain the distinction between a confidence interval for the mean response and a prediction interval for an individual response.
8. Understand what the odds ratio represents in a logistic model, and use the confidence interval to determine whether a change in condition results in a significant increase or decrease in odds.

THIS CHAPTER IN CONTEXT

Chapters 3 and 4 described the relationship between *two quantitative variables*. We used scatterplots to visualize patterns and identify linear relationships in particular. We obtained the equation of the least-square regression line to model linear relationships. In Chapter 23 we used statistical inference procedures to ask whether the linear pattern in the sample data would hold for the entire population and to estimate parameters such as the regression slope and the mean response.

In this chapter we expand the concept of simple linear regression to include multiple explanatory variables used to describe and predict one response variable. We use inference procedures to determine which of these explanatory variables, alone or in sets, has a statistically significant effect on the response variable. Some of these statistical procedures are variants of the *t* procedures from Chapter 17 and ANOVA from Chapter 24. In fact, in this chapter we find out how a categorical variable can be coded with zeros and ones to allow it into a linear regression model, instead of being used as a factor as in ANOVA.

A categorical response variable can also be coded with zeros and ones for the purpose of regression, as we describe here in the context of logistic regression. Logistic regression relies on the concepts of odds and odds ratios, which were first examined in Chapters 9 and 20, respectively.

CHECK YOUR SKILLS

Many exercise bikes, elliptical trainers, and treadmills display basic information like distance, speed, Calories burned per hour (or total Calories), and duration of the workout. The data in Table 28.6 show the treadmill display's claimed Calories per hour by speed (miles per hour) for a 175-pound male using a Cybex treadmill at inclines of 0%, 2%, and 4%.

The relationship between speed and Calories is different for walking and running, so we need an indicator for slow/fast. The variables created from Table 28.6 are

Calories=Calories burned per hourMPH=speed of the treadmillIncline=the incline percent(0,2,or 4)Ind_slow=1 for MPH≤3 andInd_slow=0 for MPH>3.0

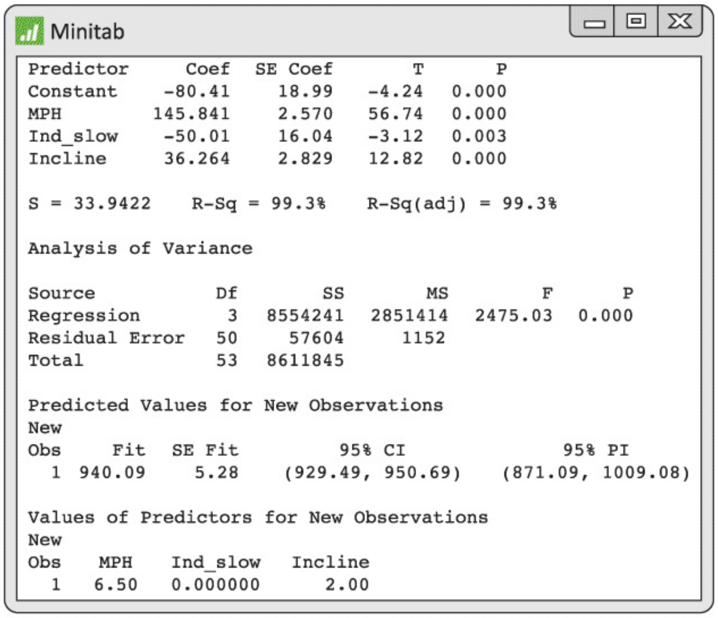
$$\begin{aligned}
&Calories = Calories\ burned\ per\ hour\\
&MPH = speed\ of\ the\ treadmill\\
&Incline = the\ incline\ percent\ (0, 2, or\ 4)\\
&Ind\_slow = 1\ for\ MPH \leq 3\ and\\
&Ind\_slow = 0\ for\ MPH > 3.0
\end{aligned}$$

TABLE 28.6 Cybex treadmill display's claimed Calories per hour by speed and incline for a 175-lb man

	Incline (%)		
MPH	**0**	**2**	**4**
1.5	174	207	240
2.0	205	249	294
2.5	236	291	347
3.0	267	333	400
3.5	372	436	503
4.0	482	542	607
4.5	592	649	709
5.0	701	756	812
5.5	763	824	885
6.0	825	892	959

6.5	887	960	1032
7.0	949	1027	1105
7.5	1011	1094	1178
8.0	1073	1163	1252
8.5	1135	1230	1325
9.0	1197	1298	1398
9.5	1259	1365	1470
10.0	1321	1433	1544

Here is part of the Minitab output from fitting a multiple regression model to predict Calories from MPH, Ind_slow, and Incline for the Cybex treadmill:



```
Predictor      Coef   SE Coef       T      P
Constant     -80.41     18.99   -4.24  0.000
MPH         145.841     2.570   56.74  0.000
Ind_slow     -50.01     16.04   -3.12  0.003
Incline      36.264     2.829   12.82  0.000

S = 33.9422   R-Sq = 99.3%   R-Sq(adj) = 99.3%

Analysis of Variance

Source          Df       SS       MS        F      P
Regression       3  8554241  2851414  2475.03  0.000
Residual Error  50    57604     1152
Total           53  8611845

Predicted Values for New Observations
New
Obs    Fit  SE Fit         95% CI               95% PI
  1 940.09    5.28  (929.49, 950.69)  (871.09, 1009.08)

Values of Predictors for New Observations
New
Obs    MPH  Ind_slow  Incline
  1   6.50  0.000000     2.00
```

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

Exercises 28.19 to 28.26 are based on this output.

28.19 The equation for predicting Calories from these explanatory variables is

a. Calories=-80.41+145.84MPH-50.01Ind_slow+36.26Incline.

$Calories = -80.41 + 145.84 MPH - 50.01 Ind\_slow + 36.26 Incline.$

b. Calories=-4.24+56.74MPH-3.12Ind_slow+12.82Incline.

$Calories = -4.24 + 56.74 MPH - 3.12 Ind\_slow + 12.82 Incline.$

c. Calories=18.99+2.57MPH+16.04Ind_slow+2.83Incline.
$Calories = 18.99 + 2.57MPH + 16.04Ind\_slow + 2.83Incline.$

28.20 The regression standard error for these data is

a. 0.993.
b. 33.94.
c. 1152.

28.21 To predict Calories when walking with no incline, use the line

a. -80.41+145.84MPH.$-80.41 + 145.84MPH.$
b. (-80.41-50.01)+145.84MPH.$(-80.41 - 50.01) + 145.84MPH.$
c. [-80.41+(2×36.26)]+145.84MPH.$[-80.41 + (2 \times 36.26)] + 145.84MPH.$

28.22 To predict Calories when running with no incline, use the line

a. -80.41+145.84MPH.$-80.41 + 145.84MPH.$
b. (-80.41-50.01)+145.84MPH.$(-80.41 - 50.01) + 145.84MPH.$
c. [-80.41+(2×36.26)]+145.84MPH.$[-80.41 + (2 \times 36.26)] + 145.84MPH.$

28.23 To predict Calories when running on a 2% incline, use the line

a. -80.41+145.84MPH.$-80.41 + 145.84MPH.$
b. (-80.41-50.01)+145.84MPH.$(-80.41 - 50.01) + 145.84MPH.$
c. [-80.41+(2×36.26)]+145.84MPH.$[-80.41 + (2 \times 36.26)] + 145.84MPH.$

28.24 Is there significant evidence that more Calories are burned for higher speeds? To answer this question, you test the hypotheses

a. H0: β0=0 versus Ha: β0>0.$H_0 : \beta_0 = 0 \text{ versus } H_a : \beta_0 > 0.$
b. H0: β1=0 versus Ha: β1>0.$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 > 0.$
c. H0: β1=0 versus Ha: β1≠0.$H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0.$

28.25 Confidence intervals and tests for these data use the *t* distribution with degrees of freedom

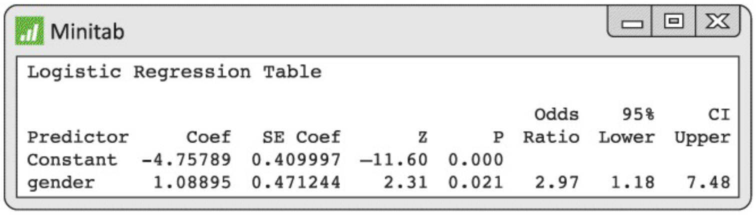
a. 3.
b. 50.
c. 53.

28.26 Orlando, a 175-pound man, plans to run 6.5 mph for 1 hour on a 2%

incline. He can be 95% confident that he will burn between

a. 871 and 1009 Cal.
b. 929 and 950 Cal.
c. 906 and 974 Cal.

Physical and hormonal factors contribute to bone weakness in aging women. A study examined the number of bone fractures among a random sample of 1469 elderly men and women. We set the response variable fracture to be equal to 1 for one or more fractures and 0 for no fracture in the previous three years. In this context, p is the probability of having a bone fracture in the past three years for an elderly person.[17] *Below is the Minitab output for the logistic regression of fracture as a function of gender (0 for men and 1 for women). Exercises 28.27 and 28.28 are based on this output.*



```
Logistic Regression Table

                                                   Odds    95%     CI
Predictor      Coef    SE Coef        Z      P    Ratio  Lower  Upper
Constant   -4.75789   0.409997   -11.60  0.000
gender      1.08895   0.471244     2.31  0.021    2.97   1.18   7.48
```

Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

28.27 The estimated logistic regression equation is

a. L=-4.758+1.089gender.$L = -4.758 + 1.089 gender$.
b. L=1.089+0.471gender.$L = 1.089 + 0.471 gender$.
c. p=-4.758+1.089gender.$p = -4.758 + 1.089 gender$.

28.28 Which of the following statements is true?

a. The odds of a fracture are not significantly different for men and for women.
b. The odds of a fracture are significantly higher for women than for men, about 3 to 7.5 times with 95% confidence.
c. The odds of a fracture are significantly higher for women than for men, about 1.2 to 7.5 times with 95% confidence.

CHAPTER 28 EXERCISES

28.29 Children's feet. A mother was told by a shoe salesman that girls' shoes are narrower than boys' shoes because girls have narrower feet. Table 28.7 shows foot measurements (in cm) collected on 39 fourth-graders.[18] Run a simple t test to determine whether there is a significant difference in foot width between boys and girls in the fourth grade. Does your analysis support the salesman's assertion? Also use software to create a regression model of *width* using only *sex_B*. Because boys and girls have similar variability in this data set, you should find that both analyses give very *similar* P-values.

TABLE 28.7 Foot width and length of fourth-graders (in cm)

Width	Length	Sex_B	Width	Length	Sex_B
8.4	22.4	1	9.2	22.0	1
8.8	23.4	1	8.6	22.4	1
9.7	22.5	1	8.3	22.0	0
9.8	23.2	1	9.0	22.5	0
8.9	23.1	1	8.1	22.2	0
9.7	23.7	1	9.4	25.1	1
9.6	24.1	1	9.5	24.1	0
8.8	21.0	0	9.5	23.5	1
9.3	21.6	0	8.9	22.2	1
8.8	20.9	1	9.3	21.9	1
9.8	25.5	1	9.3	22.0	0
8.9	22.8	1	8.6	20.5	0
9.1	24.1	1	8.6	22.5	0
9.8	25.0	1	9.0	21.6	1
9.3	24.0	0	8.6	22.7	1
7.9	21.7	0	8.5	20.9	0
8.7	22.0	0	9.0	24.0	0
8.8	22.7	0	7.9	19.6	0
9.0	24.7	0	8.8	22.6	0

9.5	23.5	0			

28.30 Children’s feet, continued. Go back to Exercise 28.29. Feet are always longer than they are wide, and we can expect that foot width is somehow associated with foot length.

a. Make a scatterplot of width (*y*) against length and describe the relationship. Use software to determine whether this relationship is statistically significant.
b. Now use software to obtain a multiple linear regression model of *width* that includes both *length* and *sex_B* as explanatory variables. Create a scatterplot using different symbols for boys and girls. Discuss the results shown in the ANOVA table and individual *t* tests in the context of your scatterplot. Check the conditions for inference. Does this analysis support the salesman’s assertion?
c. Contrast the conclusions you reached here and in Exercise 28.29. The conclusions are somewhat different. Explain why. This illustrates why regression coefficients must be interpreted in the context of the specific model used.

28.31 Caterpillars. Scientists have long been interested in the question of how body mass determines characteristics such as metabolic rate. Tobacco hornworm caterpillars (*Manduca sexta*) were chosen to study this relationship because they maintain their overall shape throughout the five stages of larval development. For theoretical reasons (explored in Exercises 28.44 and 28.45), researchers were interested more specifically in the relationship between the log of metabolic rate $y=log(MR)$$y = \log(MR)$ and the log of body mass $x1=log(BM)$.$x_1 = \log(BM)$. The full data set for caterpillars at the fourth and fifth stages of development is available in the file *ex28-31.dat* available on the companion website.[19] What is the relationship between the log of metabolic rate and the log of body mass for tobacco hornworm caterpillars? Is the relationship the same for the two different stages? Use the four-step process as in Examples 28.1, 28.12, and 28.13. Make sure to check the conditions for inference and to include a statistical analysis.

28.32 Correlated explanatory variables. Suppose that $x1=2x2-4$,

$x_1 = 2x_2 - 4$, so that x_1 and x_2 are positively correlated. Let y=3x2+4, $y = 3x_2 + 4$, so that y and x_2 are positively correlated.

a. Use the relationship between x_1 and x_2 to find the linear relationship between y and x_1. Are y and x_1 positively correlated?
b. Add the equations x1=2x2-4$x_1 = 2x_2 - 4$ and y=3x2+4$y = 3x_2 + 4$ and solve for y to obtain an equation relating y to both x_1 and x_2. Are the coefficients of both x's positive? Combining explanatory variables that are correlated can produce surprising results.

28.33 Body fat for men. You are interested in predicting the amount of body fat y of a man using the explanatory variables waist size x_1 and height x_2.

a. Do you think body fat y and waist size x_1 are positively correlated? Explain.
b. For a fixed waist size, height x_2 is negatively correlated with body fat y. Explain why.
c. The slope of the simple linear regression line for predicting body fat from height for a sample of men is almost 0, say, 0.13. Knowing a man's height does not tell you much about his body fat. Do you think this parameter estimate would become negative if a multiple regression model with height x_2 and waist size x_1 was used to predict body fat? Explain.

28.34 Metabolic rate and body mass. Metabolic rate, the rate at which the body consumes energy, is important in studies of weight gain, dieting, and exercise. The table below gives data on the lean body mass and resting metabolic rate for 12 women and 7 men who are subjects in a study of dieting. Lean body mass, given in kilograms, is a person's weight leaving out all fat. Metabolic rate is measured in kilocalories (Cal) burned per 24 hours, the same calories used to describe the energy content of foods. The researchers believe that lean body mass is an important influence on metabolic rate.

Subject	Sex	Mass	Rate	Subject	Sex	Mass	Rate

1	M	62.0	1792
2	M	62.9	1666
3	F	36.1	995
4	F	54.6	1425
5	F	48.5	1396
6	F	42.0	1418
7	M	47.4	1362
8	F	50.6	1502
9	F	42.0	1256
10	M	48.7	1614
11	F	40.3	1189
12	F	33.1	913
13	M	51.9	1460
14	F	42.4	1124
15	F	34.5	1052
16	F	51.1	1347
17	F	41.2	1204
18	M	51.9	1867
19	M	46.9	1439

a. Make a scatterplot of the data, using different symbols or colors for men and women. Summarize what you see in the plot.

b. Use the model with two regression lines to predict metabolic rate from lean body mass for the different genders. Summarize the results.

c. The parameter associated with the interaction term is often used to decide if a model with parallel regression lines can be used. Test the hypothesis that this parameter is equal to zero, and comment on whether you would be willing to use the more restrictive model with parallel regression lines for these data.

28.35 World record running times. Table 28.8 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.

a. Make a scatterplot of world record time against year, using separate symbols for men and women. Describe the pattern for each sex. Then compare the progress of men and women.

b. Fit the model with two regression lines, one for women and one for men, and identify the estimated regression lines.

c. Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?

TABLE 28.8 World record running times for men and women

Men				Women	
Record year	Time (seconds)	Record year	Time (seconds)	Record year	Time (seconds)
1912	1880.8	1962	1698.2	1967	2286.4
1921	1840.2	1963	1695.6	1970	2130.5
1924	1835.4	1965	1659.3	1975	2100.4
1924	1823.2	1972	1658.4	1975	2041.4
1924	1806.2	1973	1650.8	1977	1995.1
1937	1805.6	1977	1650.5	1979	1972.5
1938	1802.0	1978	1642.4	1981	1950.8
1939	1792.6	1984	1633.8	1981	1937.2
1944	1775.4	1989	1628.2	1982	1895.2
1949	1768.2	1993	1627.9	1983	1895.0
1949	1767.2	1993	1618.4	1983	1887.6
1949	1761.2	1994	1612.2	1984	1873.8
1950	1742.6	1995	1603.5	1985	1859.4
1953	1741.6	1996	1598.1	1986	1813.7
1954	1734.2	1997	1591.3	1993	1771.8
1956	1722.8	1997	1587.8		
1956	1710.4	1998	1582.7		
1960	1698.8	2004	1580.3		

28.36 World record running times, continued. Exercise 28.35 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and women.

a. Provide the ANOVA table for the regression model with two regression lines, one for men and one for women.
b. Are all the individual coefficients significantly different from zero? Set up the appropriate hypotheses, identify the test statistics and *P*-values, and make conclusions in the context of the problem.

28.37 Mycorrhizal colonies and plant nutrition. Mycorrhizal fungi are present in the roots of many plants. This is a symbiotic relationship in which the plant supplies nutrition to the fungus and the fungus helps the plant absorb nutrients from the soil. An experiment compared the effects of adding various amounts of nitrogen fertilizer (in kilograms per hectare) to two genotypes of tomato plants, a wild-type variety with mycorrhizal colonies and a mutant variety without the colonies. The percent of phosphorus in the plant at harvest time was then assessed, and the results are shown in Table 28.9.[20]

a. Produce a scatterplot of the amount of phosphorus in the plant against the amount of fertilizer provided (nitrogen), and use different symbols for the two plant genotypes. Does the graph suggest that a multiple linear regression might be appropriate? Explain.

b. Use software to obtain the estimated multiple linear regression equation when the two explanatory variables nitrogen and genotype are included. Create a residual plot for this model. Are the conditions for multiple linear regression satisfied? Explain.

c. Create a new variable called “Interaction” by multiplying the explanatory variables nitrogen and genotype. Add this new variable to your regression model. Provide the estimated multiple linear regression equation. Create a residual plot for this new model and discuss whether the conditions for multiple linear regression are met.

d. Does the ANOVA table for the model with the interaction term indicate that at least one of the explanatory variables is helpful in predicting the amount of phosphorus in the plant? Do the individual *t* tests indicate that all coefficients are significantly different from zero? Explain.

TABLE 28.9 Percent of phosphorus in tomato plants

Phosphorus	Nitrogen	Genotype	Phosphorus	Nitrogen	Genotype
0.29	0	0	0.64	0	1
0.25	0	0	0.54	0	1
0.27	0	0	0.53	0	1
0.24	0	0	0.52	0	1
0.24	0	0	0.41	0	1
0.2	0	0	0.43	0	1
0.21	28	0	0.41	28	1
0.24	28	0	0.37	28	1

0.21	28	0	0.5	28	1
0.22	28	0	0.43	28	1
0.19	28	0	0.39	28	1
0.17	28	0	0.44	28	1
0.18	160	0	0.34	160	1
0.20	160	0	0.31	160	1
0.19	160	0	0.36	160	1
0.19	160	0	0.37	160	1
0.16	160	0	0.26	160	1
0.17	160	0	0.27	160	1

28.38 Mycorrhizal colonies and plant nutrition, continued. In Chapter 26, we ran a two-way ANOVA on this same data set. Go back to Example 26.12 and read it carefully. Explain why we can run either a two-way ANOVA or a multiple linear regression analysis on this data set. Compare your conclusions from Exercise 28.37 with those reached in Example 26.12. What are the advantages and disadvantages of either method?

28.39 Reaction time. A learning system includes a test of skill in using the computer's mouse. The software displays a circle at a random location on the computer screen. The subject clicks in the circle with the mouse as quickly as possible. A new circle appears as soon as the subject clicks the old one. Table 28.10 gives data for one subject's trials, 20 with each hand. Distance is the distance from the cursor location to the center of the new circle, in units whose actual size depends on the size of the screen. Time is the time required to click in the new circle, in milliseconds.[21]

TABLE 28.10 Reaction times (ms) in a computer game

Time	Distance	Hand	Time	Distance	Hand
115	190.70	Right	240	190.70	Left
96	138.52	Right	190	138.52	Left
110	165.08	Right	170	165.08	Left
100	126.19	Right	125	126.19	Left
111	163.19	Right	315	163.19	Left
101	305.66	Right	240	305.66	Left

111	176.15	Right	141	176.15	Left
106	162.78	Right	210	162.78	Left
96	147.87	Right	200	147.87	Left
96	271.46	Right	401	271.46	Left
95	40.25	Right	320	40.25	Left
96	24.76	Right	113	24.76	Left
96	104.80	Right	176	104.80	Left
106	136.80	Right	211	136.80	Left
100	308.60	Right	238	308.60	Left
113	279.80	Right	316	279.80	Left
123	125.51	Right	176	125.51	Left
111	329.80	Right	173	329.80	Left
95	51.66	Right	210	51.66	Left
108	201.95	Right	170	201.95	Left

a. Specify the population multiple regression model for predicting time from distance separately for each hand. Make sure you include the interaction term that is necessary to allow for the possibility of having different slopes. Explain in words what each β in your model means.

b. Use statistical software to find the estimated multiple regression equation for predicting time from distance separately for each hand. What percent of variation in the distances is explained by this multiple regression model?

c. Explain how to use the estimated multiple regression equation in part b to obtain the least-squares line for each hand. Draw these lines on a scatterplot of time versus distance.

28.40 Reaction time, continued. Go back to Exercise 28.39. Would you be willing to use your model to predict reaction times for individuals in the general population? Explain your answer. A statistical model represents only the population from which the sample was taken.

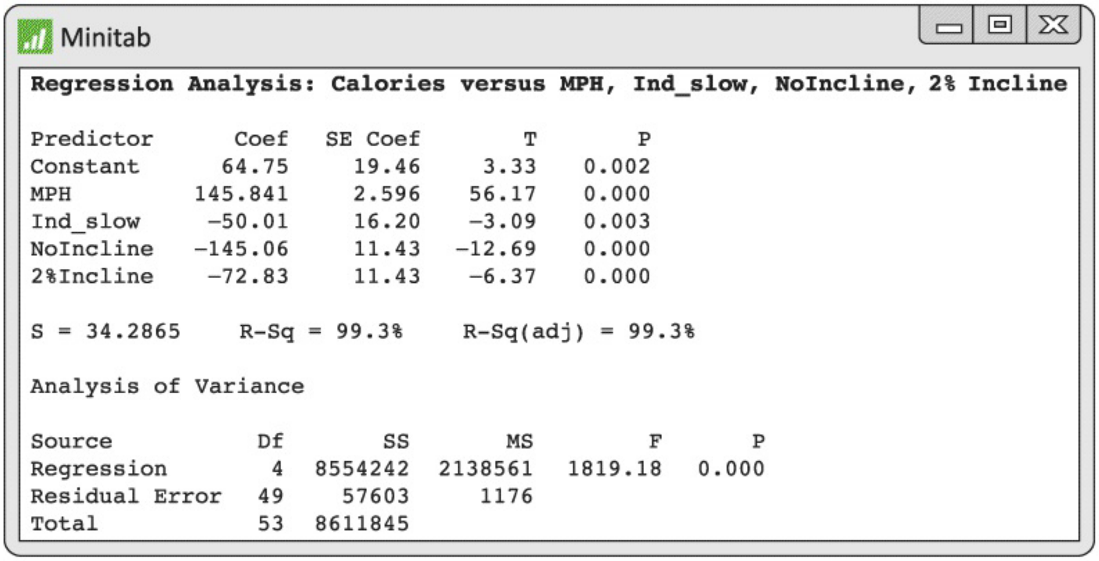
28.41 Burning calories with exercise. Many exercise bikes, elliptical trainers, and treadmills display basic information like distance, speed, Calories burned per hour (or total Calories), and duration of the workout.

Let’s take another look at the data in Table 28.6 that were used for the Check Your Skills exercises. Scatterplots show different linear relationships for each incline, one for slow speeds and another for faster speeds, so the following indicator variables were created:

Ind_slow=1 for MPH≤3 and Ind_slow=0 for MPH>3.0
NonIncline=1 for 0% incline and NonIncline=0 for other inclines
2%Incline=1 for a 2% incline and 2%Incline=0 for other inclines

$Ind\_slow = 1$ for $MPH \le 3$ and $Ind\_slow = 0$
for $MPH > 3.0$
$NonIncline = 1$ for 0% incline and $NonIncline = 0$
for other inclines
$2\%Incline = 1$ for a 2% incline and $2\%Incline = 0$
for other inclines

Figure 28.17 shows part of the Minitab output from fitting a multiple regression model to predict *Calories* from *MPH*, *Ind_slow*, *NoIncline*, and *2%Incline* for a 175-pound man using the Cybex machine.



```
Regression Analysis: Calories versus MPH, Ind_slow, NoIncline, 2% Incline

Predictor        Coef   SE Coef        T       P
Constant        64.75     19.46     3.33   0.002
MPH           145.841     2.596    56.17   0.000
Ind_slow       -50.01     16.20    -3.09   0.003
NoIncline     -145.06     11.43   -12.69   0.000
2%Incline      -72.83     11.43    -6.37   0.000

S = 34.2865     R-Sq = 99.3%     R-Sq(adj) = 99.3%

Analysis of Variance

Source           Df       SS       MS        F      P
Regression        4  8554242  2138561  1819.18  0.000
Residual Error   49    57603     1176
Total            53  8611845
```

Figure 28.17
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.17 Minitab output for Exercise 28.42.

a. Use the Minitab output to estimate each parameter in this multiple regression model for predicting Calories burned with the Cybex machine. Don’t forget to estimate *σ*.

b. How many separate lines are fitted with this model? Do the lines all have the same slope? Identify each fitted line.

c. Do you think that this model provides a good fit for these data? Explain.

d. Is there significant evidence that more calories are burned for higher speeds? State the hypotheses, identify the test statistic and *P*-value, and provide a conclusion in the context of this question.

28.42 Children's perception of reading difficulty. Table 28.11 contains measured and self-estimated reading ability data for 60 fifth-grade students randomly sampled from one elementary school.[22] The variables are:

Variable	Description
OBS	Observation number for each individual
SEX	Gender of the individual
LSS	Median grade level of student's selection of "best for me to read" (8 repetitions, each with four choices at grades 3, 5, 7, and 9 level)
IQ	IQ score
READ	Score on reading subtest of the Metropolitan Achievement Test
EST	Student's own estimate of his or her reading ability, scale 1 to 5 (1 = low)

TABLE 28.11 Measured and self-estimated reading ability data for 60 fifth-grade students randomly sampled from one elementary school

OBS	SEX	LSS	IQ	READ	EST	OBS	SEX	LSS	IQ	READ	EST
1	F	5.00	145	98	4	31	M	7.00	106	55	4
2	F	8.00	139	98	5	32	M	6.00	124	70	4
3	M	6.00	126	90	5	33	M	8.00	115	82	5
4	F	5.33	122	98	5	34	M	8.40	133	94	5
5	F	5.60	125	55	4	35	F	5.00	116	75	4
6	M	9.00	130	95	3	36	F	6.66	102	80	3
7	M	5.00	96	50	4	37	F	5.00	127	85	4
8	M	4.66	110	50	4	38	M	6.50	117	88	5
9	F	4.66	118	75	4	39	F	5.00	109	70	3
10	F	8.20	118	75	5	40	M	5.50	137	80	4
11	M	4.66	101	65	4	41	M	6.66	117	55	4
12	M	7.50	142	68	5	42	M	6.00	90	65	2
13	F	5.00	134	80	4	43	F	4.00	103	30	1

14	M	7.00	124	10	4
15	M	6.00	112	67	4
16	M	6.00	109	83	3
17	F	5.33	134	90	4
18	M	6.00	113	90	5
19	M	6.00	81	55	3
20	F	6.00	113	83	4
21	M	6.00	123	65	4
22	F	4.66	94	25	3
23	M	4.50	100	45	3
24	F	6.00	136	97	4
25	M	5.33	109	75	4
26	F	3.60	131	70	4
27	M	4.00	117	23	3
28	M	6.40	110	45	3
29	F	6.00	127	70	2
30	F	6.00	124	85	5
44	F	5.50	114	74	5
45	M	5.00	139	80	5
46	M	6.66	101	70	2
47	F	8.33	122	60	4
48	F	6.50	105	45	2
49	F	4.00	97	45	1
50	M	5.50	89	55	4
51	M	5.00	102	30	2
52	F	4.00	108	10	4
53	M	4.66	110	40	1
54	M	5.33	128	65	1
55	M	5.20	114	15	2
56	M	4.00	112	62	2
57	F	3.60	114	98	4
58	M	6.00	102	52	2
59	F	4.60	82	23	1
60	M	5.33	101	35	2

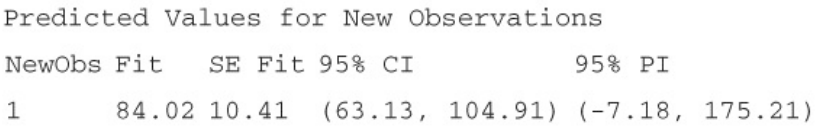
a. Is the relationship between measured (*READ*) and self-estimated (*EST*) reading ability the same for both boys and girls? Create an indicator variable for gender and fit an appropriate multiple regression model to answer the question.

b. Fit a multiple regression model for predicting *IQ* from the explanatory variables *LSS*, *READ*, and *EST*. Are you happy with the fit of this model? Explain.

c. Use residual plots to check the appropriate conditions for your model.

d. Only two of the three explanatory variables in your model in part b have parameters that are significantly different from zero according to the individual *t* tests. Drop the explanatory variable that is not significant, and add the interaction term for the two remaining explanatory variables. Are you surprised by the results from fitting this new model? Explain what happened to the individual *t* tests for the two explanatory variables.

28.43 Fish sizes. In Exercise 28.13 you built a multiple regression model with three explanatory variables (length, width, and their interaction) to predict the weight of a perch. Minitab provides the following information for a perch with length 21 cm and width 2.8 cm (which creates an interaction term of 21×2.8-58.8$21 \times 2.8 - 58.8$):



```
Predicted Values for New Observations
NewObs Fit    SE Fit 95% CI           95% PI
1      84.02 10.41  (63.13, 104.91) (-7.18, 175.21)
```

Interpret both intervals in the context of the data set.

28.44 More on caterpillars. Exercise 28.31 described the relationship between basal metabolic rate (*MR*) and body mass (*BM*) in tobacco hornworm caterpillars, or, more specifically, the relationship between response variable y=log(MR)$y = \log(MR)$ and explanatory variablex1=log(BM). `Unexpected text node: '.'` Both experimental and theoretical research have suggested the general relationship MR=α(BM)β. $MR = \alpha(BM)^{\beta}$.[23] However, there is still considerable debate on whether the scaling exponent is β=2⁄3$\beta = 2/3$ or β=3⁄4.$\beta = 3/4$.

a. Use software to estimate *α* and *β* in the general relationship MR-α(BM)β,$MR - \alpha(BM)^{\beta}$, which is the same as μlog(MR)=log(α)+βlog(BM).$\mu_{\log(\mathrm{MR})} = \log(\alpha) + \beta \log(BM)$. The predicted model is y^=a+blog(BM),$\hat{y} = a + b \log(\mathrm{BM})$, so that b_0 estimates log(α) $\log(\alpha)$ and b_1 estimates *β* in the original model. What percent of variation in log(*MR*) is explained by using linear regression with the explanatory variable log(*MR*)?
b. Find a 95% confidence interval for the slope parameter *β*. Are the values β=2⁄3$\beta = 2/3$ and β=3⁄4$\beta = 3/4$ contained in your confidence interval?

28.45 More on caterpillars, continued. Go back to Exercise 28.44. This time, use the multiple regression model with two explanatory variables: the continuous variable x1=log(BM)$x_1 = \log(BM)$ and an indicator variable reflecting instar stage (x_2 = 1 for instar 5, x_2 = 0 for instar 4). This should reflect what you did in Exercise 28.31.

a. Find a 95% confidence interval for the slope parameter β for caterpillars during stage 4.
b. If you were asked to report a confidence interval for the slope parameter β for caterpillars during stage 5, would you report the same interval that you calculated in part a? Explain why or why not.
c. Are the values $\beta=2/3$ $\beta = 2/3$ and $\beta=3/4$ $\beta = 3/4$ contained in your confidence interval from part a?
d. How does your confidence interval in part a compare with the confidence interval you computed in part b of Exercise 28.44?

28.46 Recurring bladder tumors, continued. Go back to Exercises 28.15 and 28.16. You already computed the estimated value of *p* when *x* is 0, 2, or 6. Complete your analysis for the other values of *x* (1, 3, 4, 5, 7, 8). Use your results to create a plot that shows how *p* changes as a function of *x*. Interpret this plot.

28.47 FDA warning on insomnia prescriptions. In a January 2013 safety announcement, the Food and Drug Administration warned drug manufacturers that the recommended dose of the widely prescribed insomnia drug zolpidem should be halved for women. The warning came after pharmacokinetic trials found that 38 of 250 women and 8 of 250 men who had taken the recommended dose of zolpidem still had enough drug in their blood 8 hours later to cause impairment.

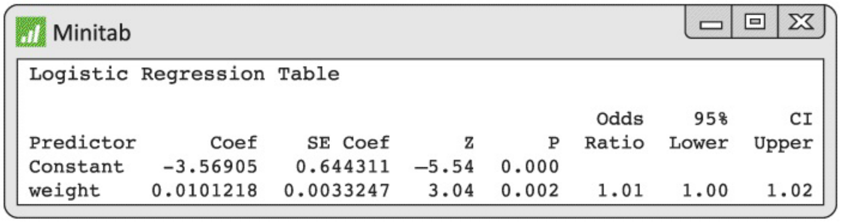
a. What are the odds of impairment 8 hours after dosing for women and for men? Use this information to compute the equation of the logistic regression model if the indicator variable *x* equals 0 for men and 1 for women.
b. Minitab provides the standard error of the slope, SEb1=0.4002. $\text{SE}_{b1} = 0.4002.$ Use this information to compute a 95% confidence interval for the slope of the logistic regression model. Then compute a 95% confidence interval for the odds ratio for impairment 8 hours after dosing. Interpret your results in the context of the FDA warning.

28.48 Treating AIDS. The drug AZT was the first drug that seemed effective in delaying the onset of AIDS in people infected with HIV. A study randomly assigned 435 HIV-positive volunteers who had not yet developed AIDS to take 500 mg of AZT each day (AZT=1$AZT = 1$) and another 435 to take a placebo (AZT=0$AZT = 0$). At the end of the study, 38 of the placebo subjects and 17 of the AZT subjects had developed AIDS.

a. What are the odds of developing AIDS in each of the two groups? Use this information to compute the equation of the logistic regression model using the indicator variable *AZT*.
b. Minitab provides the standard error of the slope, SEb1=0.3001. $\mathrm{SE}_{b1} = 0.3001$. Use this information to compute a 95% confidence interval for the slope of the logistic regression model. Then compute a 95% confidence interval for the odds ratio for developing AIDS. Do the data support the claim that taking AZT lowers the odds of developing AIDS in HIV-positive patients? Explain.

A study screened a random sample of adult subjects for adult-onset, type 2 diabetes with the objective of identifying some of the risk factors associated with the disease.The file ex28-49.dat on the companion website contains data about diabetes status, weight (in pounds), waist circumference (in inches), and cholesterol ratio (ratio of total cholesterol to HDL cholesterol in blood) for all 386 subjects.[24] *Fifty-nine of the 386 were identified as having type 2 diabetes (diabetes = 1). Exercises 28.49 to 28.51 are based on this study.*

28.49 Type 2 diabetes. Adult-onset, or type 2, diabetes has been associated with obesity. The Minitab output in Figure 28.18 shows the results of a logistic regression analysis with response variable *diabetes* and explanatory variable *weight*. Give the equation of the logistic regression model. Is the coefficient for the slope significantly different from zero? Does a higher weight appear to increase the odds of having type 2 diabetes?



```
Logistic Regression Table

                                                    Odds     95%     CI
Predictor        Coef     SE Coef       Z       P  Ratio   Lower  Upper
Constant     -3.56905    0.644311   -5.54   0.000
weight      0.0101218   0.0033247    3.04   0.002   1.01    1.00   1.02
```

Figure 28.18
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company

FIGURE 28.18 Minitab output for Exercises 28.49 to 28.51.

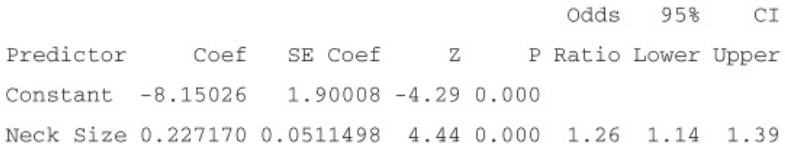
28.50 Type 2 diabetes, continued. Use software to obtain the logistic regression analysis with response variable *diabetes* and explanatory variable *weight*. Make sure that you obtain the same results as those shown in Exercise 28.49, up to rounding error.

a. Research suggests that where body fat is stored is an important factor in predicting diabetes. Use software to run a model with both *weight* and *waist* as explanatory variables. What is the equation of this new logistic regression model? Do both variables contribute significantly to the model? Explain how adding the variable *waist* could change the relative contribution of the variable *weight*.
b. Use software to fit a logistic regression model to predict *diabetes* from *waist*. Interpret the output in the context of the study. Which of the three models described in this exercise would you favor and why?

28.51 Type 2 diabetes, continued. Among the list of potential risk factors for type 2 diabetes is a person’s cholesterol ratio (*cholratio*), the ratio of total cholesterol to good, or HDL, cholesterol. Use software to fit a logistic regression model to predict *diabetes* from both *waist* and *cholratio* as explanatory variables. What is the equation of this new logistic regression model? Are both coefficients significant? Explain in your own words what the two corresponding odds ratios mean in the context of this study.

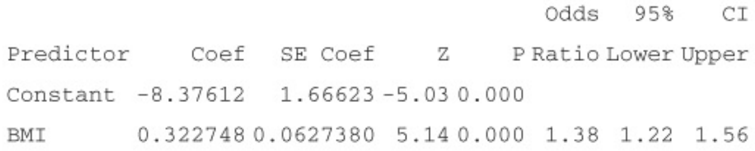
28.52 Sleep apnea. Sleep apnea is a potentially dangerous condition in which patients cease to breathe for at least 10 seconds at a time while asleep. Researchers collected data on 133 individuals in order to find risk factors for sleep apnea.

a. Neck size (measured in centimeters) may predict the risk of sleep apnea because a larger neck size might constrict the airway. Below is the software output for the logistic regression model explaining sleep apnea (y=1$y = 1$ for sleep apnea, 0 otherwise) as a function of neck size. Is neck size significant in predicting sleep apnea? Explain.



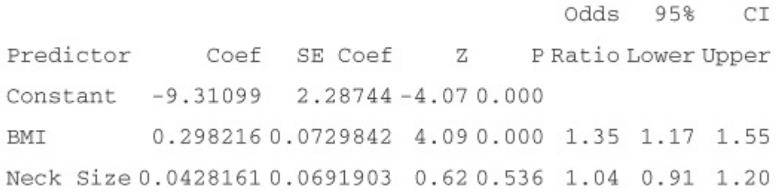
```
                                                   Odds   95%    CI
Predictor      Coef     SE Coef     Z      P Ratio Lower Upper
Constant  -8.15026   1.90008 -4.29 0.000
Neck Size 0.227170 0.0511498  4.44 0.000  1.26  1.14  1.39
```

b. Being overweight has been linked to sleep apnea. Below is the software output for the logistic regression model explaining sleep apnea as a function of an individual’s body mass index (BMI, measured in kg/m^2). Is BMI significant in predicting sleep apnea? Explain.



```
                                                  Odds   95%    CI
Predictor      Coef    SE Coef     Z      P Ratio Lower Upper
Constant  -8.37612   1.66623 -5.03 0.000
BMI        0.322748 0.0627380  5.14 0.000  1.38  1.22  1.56
```

c. Based on the results obtained in a and b, we also ran a logistic regression model using both neck size and BMI as explanatory variables. The software output is given below. Are both variables still significant when the other is taken into consideration? What conclusions do you reach?



```
                                                   Odds   95%    CI
Predictor       Coef    SE Coef     Z      P Ratio Lower Upper
Constant   -9.31099   2.28744 -4.07 0.000
BMI         0.298216 0.0729842  4.09 0.000  1.35  1.17  1.55
Neck Size 0.0428161 0.0691903  0.62 0.536  1.04  0.91  1.20
```

28.53 Student achievement and self-concept. In order to determine if student achievement is related to self-concept, as measured by the Piers-Harris Children’s Self-Concept Scale, data were collected on 78 seventh-grade students from a rural midwestern school. The large data file *Large.SelfConcept* available on the companion website contains data on the following variables for all 78 students:[25]

Variable	Description
OBS	Observation number (n = 78, some

	gaps in numbers)
GPA	GPA from school records
IQ	Test score from school records
AGE	Age in years, self-reported
GENDER	1 = F, 2 = M, self-reported
RAW	Raw score on Piers-Harris Childrens' Self-Concept Scale
C1	Cluster 1 within self-concept: behavior
C2	Cluster 2: school status
C3	Cluster 3: physical
C4	Cluster 4: anxiety
C5	Cluster 5: popularity
C6	Cluster 6: happiness

We will investigate the relationship between *GPA* and only three of the explanatory variables:

- *IQ*, the student's score on a standard IQ test
- *C*2, the student's self-assessment of his or her school status
- *C*5, the student's self-assessment of his or her popularity

Use statistical software to analyze the relationship between students' *GPA* and their *IQ*, self-assessed school status (*C*2), and self-assessed popularity (*C*5).

a. One observation is an extreme outlier when all three explanatory variables are used. Which observation is this? Give the observation number and explain how you found it using regression output. Find this observation in the data list. What is unusual about it?
b. Software packages often identify unusual or influential observations. Have any observations been identified as unusual or influential? If so, identify these points on a scatterplot of *GPA* versus *IQ*.
c. *C*2 (school status) is the aspect of self-concept most highly correlated to *GPA*. If we carried out the simple linear regression of *GPA* on *C*2, what percent of the variation in students' GPAs would be explained by the straight-line relationship between *GPA* and *C*2?

d. You know that *IQ* is associated with *GPA*, and you are not studying that relationship. Because *C*2 and *IQ* are positively correlated (r=0.547 $r = 0.547$), a significant relationship between *C*2 and *GPA* might occur just because *C*2 can “stand in” for *IQ*. Does *C*2 still contribute significantly to explaining *GPA* after we have allowed for the relationship between *GPA* and *IQ*? (Give a test statistic, its *P*-value, and your conclusion.)

e. A new student in this class has *IQ* 115 and *C*2 score 14. What do you predict this student’s *GPA* to be? (Just give a point prediction, not an interval.)

28.54 Growth of pine trees. The Department of Biology at Kenyon College conducted an experiment to study the growth of pine trees at a site located just south of Gambier on a hill overlooking the Kokosing River. In April 1990, student and faculty volunteers planted 1000 white pine (*Pinus strobus*) seedlings at the Brown Family Environmental Center. These seedlings were planted in two grids, distinguished by 10- and 15-foot spacings between the seedlings. A subset of the data collected by students at Kenyon College is available in the *Large.PineTrees* file on the companion website.[26] A description of the variables is provided below.

Variable	Description
Row	Row number in pine plantation
Col	Column number in pine plantation
Hgt90	Tree height at the time of planting (cm)
Hgt96	Tree height in September 1996 (cm)
Diam96	Tree trunk diameter in September 1996 (cm)
Grow96	Leader growth during 1996 (cm)
Hgt97	Tree height in September 1997 (cm)
Diam97	Tree trunk diameter in September 1997 (cm)
Spread97	Widest lateral spread in September 1997 (cm)
Needles97	Needle length in September 1997 (mm)
Deer95	Type of deer damage in September 1995: 1 = none, 2 = browsed
Deer97	Type of deer damage in September 1997:

	1 = none, 2 = browsed
Cover95	Amount of thorny cover in September 1995: 0 = none, 1 = < 1⁄3, 2 = between 1/3 and 2/3, 3 = > 2⁄3
Fert	Indicator for fertilizer: 0 = no, 1 = yes
Spacing	Distance (in feet) between trees (10 or 15)

a. Use tree height at the time of planting (*Hgt90*) and the indicator variable for fertilizer (*Fert*) to fit a multiple regression model for predicting *Hgt97*. Specify the estimated regression model and the regression standard error. Are you happy with the fit of this model? Comment on the value of R^2 and the plot of the residuals against the predicted values.

b. Construct a correlation matrix with *Hgt90*, *Hgt96*, *Diam96*, *Grow96*, *Hgt97*, *Diam97*, *Spread97*, and *Needles97*. Which variable is most strongly correlated with the response variable of interest (*Hgt97*)? Does this make sense to you?

c. Add tree height in September 1996 (*Hgt96*) to the model in part a. Does this model do a better job of predicting tree height in 1997? Explain.

d. What happened to the individual t statistic for *Hgt90* when *Hgt96* was added to the model? Explain why this change occurred.

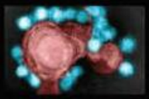
e. Fit a multiple regression model for predicting *Hgt97* based on the explanatory variables *Diam97*, *Hgt96*, and *Fert*. Summarize the results of the individual t tests. Does this model provide a better fit than the previous models? Explain by comparing the values of R^2 and s for each model.

f. Does the parameter estimate for the variable indicating whether a tree was fertilized or not have the sign you expected? Explain. (Experiments can produce surprising results!)

g. Do you think that the model in part e should be used for predicting growth in other pine seedlings? Think carefully about the conditions for inference.

The Practice of Statistics in the Life Sciences (PSLS) brings a data analysis approach to the introductory biostatistics course, with a focus on statistical reasoning and actual studies drawn from a variety of fields. The following applications are new to *PSLS*, 4e:

We consider several studies of concussion, looking at the long-term physiological effects of head trauma and examining the association between contact sports, head trauma, and chronic traumatic encephalopathy (CTE).

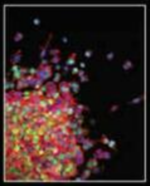


We examine the scientific evidence behind the causal link between the Zika virus and microcephaly in newborns. We also look at studies using animal models of Zika viral infection to understand its mechanism of action and find a possible treatment.

Researchers have documented the world's largest breeding colony of Magellan penguins for nearly three decades. We discuss their findings and how they addressed outlier years with extraordinary floods.

Make every problem count with SaplingPlus!

The fourth edition is supported by SaplingPlus, a powerful homework system interwoven with the content and resources from the text. SaplingPlus makes every problem a learning opportunity through targeted feedback, hints, and detailed solutions. See purchase options at www.macmillanlearning.com/psls4e.



About the Cover:

Lung cancer cells invading healthy tissue. Researchers found an association between lifetime total number of stem cell divisions and lifetime risk of cancer for various organs, suggesting that random mutations explain in part why some cancers are more common than others.

Front Cover Photo Credit: National Cancer Institute
/Winship Cancer Institute of Emory University
Back Cover Photo Credits: Top: mark@rocketclips.com/Depositphotos.
com; middle: Science Source; bottom: encrier/Depositphotos.com

w. h. freeman
Macmillan Learning
macmillanlearning.com

Contents

Front Matter

1. Cover Page
2. Title Page
3. Copyright Page
4. Brief Contents
5. Contents
6. To the Instructor: About this Book
7. Acknowledgments
8. Media and Supplements
9. To the Student: Statistical Thinking
10. About the Authors

Part I: Collecting and Exploring Data

Chapter 1 Picturing Distributions with Graphs

Individuals and variables

Identifying categorical and quantitative variables

Categorical variables: pie charts and bar graphs

Quantitative variables: histograms

Interpreting histograms

Quantitative variables: dotplots

Time plots

Discussion: (Mis)adventures in data entry

Chapter 1 Summary

This Chapter in Context

Check Your Skills

Chapter 1 Exercises

Chapter 2 Describing Quantitative Distributions with Numbers

Measures of center: median, mean

Measures of spread: percentiles, standard deviation

- Graphical displays of numerical summaries
- Spotting suspected outliers
- Discussion: Dealing with outliers
- Organizing a statistical problem
- Chapter 2 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 2 Exercises

Chapter 3 Scatterplots and Correlation
- Explanatory and response variables
- Relationship between two quantitative variables: scatterplots
- Adding categorical variables to scatterplots
- Measuring linear association: correlation
- Chapter 3 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 3 Exercises

Chapter 4 Regression
- The least-squares regression line
- Facts about least-squares regression
- Outliers and influential observations
- Working with logarithm transformations
- Cautions about correlation and regression
- Association does not imply causation
- Chapter 4 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 4 Exercises

Chapter 5 Two-Way Tables
- Marginal distributions
- Conditional distributions

- Simpson’s paradox
- Chapter 5 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 5 Exercises

Chapter 6 Samples and Observational Studies
- Observation versus experiment
- Sampling
- Sampling designs
- Sample surveys
- Cohorts and case-control studies
- Chapter 6 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 6 Exercises

Chapter 7 Designing Experiments
- Designing experiments
- Randomized comparative experiments
- Common experimental designs
- Cautions about experimentation
- Ethics in experimentation
- Discussion: The Tuskegee syphilis study
- Chapter 7 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 7 Exercises

Chapter 8 Collecting and Exploring Data: Part I Review
- Part I summary
- Comprehensive review exercises
- Large data set exercises
- Online data sources

    - EESEE case studies
- Part II: From Chance to Inference
  - Chapter 9 Essential Probability Rules
    - The idea of probability
    - Probability models
    - Probability rules
    - Discrete versus continuous probability models
    - Random variables
    - Risk and odds
    - Chapter 9 Summary
    - This Chapter in Context
    - Check Your Skills
    - Chapter 9 Exercises
  - Chapter 10 Independence and Conditional Probabilities
    - Relationships among several events
    - Conditional probability
    - General probability rules
    - Tree diagrams
    - Bayes’s theorem
    - Discussion: Making sense of conditional probabilities in diagnostic tests
    - Chapter 10 Summary
    - This Chapter in Context
    - Check Your Skills
    - Chapter 10 Exercises
  - Chapter 11 The Normal Distributions
    - Normal distributions
    - The 68–95–99.7 rule
    - The standard Normal distribution
    - Finding Normal probabilities
    - Finding percentiles

  - Using the standard Normal table
  - Normal quantile plots
  - Chapter 11 Summary
  - This Chapter in Context
  - Check Your Skills
  - Chapter 11 Exercises
- Chapter 12 Discrete Probability Distributions
  - The binomial setting and binomial distributions
  - Binomial probabilities
  - Binomial mean and standard deviation
  - The Normal approximation to binomial distributions
  - The Poisson distributions
  - Poisson probabilities
  - Chapter 12 Summary
  - This Chapter in Context
  - Check Your Skills
  - Chapter 12 Exercises
- Chapter 13 Sampling Distributions
  - Parameters and statistics
  - Statistical estimation and sampling distributions
  - The sampling distribution of $\bar{x}$
  - The central limit theorem
  - The sampling distribution of $\hat{p}$
  - The law of large numbers
  - Chapter 13 Summary
  - This Chapter in Context
  - Check Your Skills
  - Chapter 13 Exercises
- Chapter 14 Introduction to Inference
  - Statistical estimation
  - Margin of error and confidence level

Confidence intervals for the mean μ
Hypothesis testing
P-value and statistical significance
Tests for a population mean
Tests from confidence intervals
Chapter 14 Summary
This Chapter in Context
Check Your Skills
Chapter 14 Exercises
Chapter 15 Inference in Practice
Conditions for inference in practice
How confidence intervals behave
How hypothesis tests behave
Discussion: The scientific approach
Planning studies: selecting an appropriate sample size
Chapter 15 Summary
This Chapter in Context
Check Your Skills
Chapter 15 Exercises
Chapter 16 From Chance to Inference: Part II Review
Part II summary
Comprehensive review exercises
Advanced topics (optional material)
Online data sources
EESEE case studies
Part III: Statistical Inference
Chapter 17 Inference about a Population Mean
Conditions for inference
The t distributions
The one-sample t confidence interval
The one-sample t test

- Matched pairs t procedures
- Robustness of t procedures
- Chapter 17 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 17 Exercises

Chapter 18 Comparing Two Means

- Comparing two population means
- Two-sample t procedures
- Robustness again
- Avoid the pooled two-sample t procedures
- Avoid inference about standard deviations
- Chapter 18 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 18 Exercises

Chapter 19 Inference about a Population Proportion

- The sample proportion $\hat{p}$
- Large-sample confidence intervals for a proportion
- Accurate confidence intervals for a proportion
- Choosing the sample size
- Hypothesis tests for a proportion
- Chapter 19 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 19 Exercises

Chapter 20 Comparing Two Proportions

- Two-sample problems: proportions
- The sampling distribution of a difference between proportions
- Large-sample confidence intervals for comparing proportions
- Accurate confidence intervals for comparing proportions

- Hypothesis tests for comparing proportions
- Relative risk and odds ratio
- Discussion: Assessing and understanding health risks
- Chapter 20 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 20 Exercises

Chapter 21 The Chi-Square Test for Goodness of Fit
- Hypotheses for goodness of fit
- Expected counts and chi-square statistic
- The chi-square test for goodness of fit
- Interpreting significant chi-square results
- Conditions for the chi-square test
- Chapter 21 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 21 Exercises

Chapter 22 The Chi-Square Test for Two-Way Tables
- Two-way tables
- Hypotheses for two-way tables of counts
- Expected counts and chi-square statistic
- The chi-square test
- Conditions for the chi-square test
- The chi-square test and the two-sample z test
- Chapter 22 Summary
- This Chapter in Context
- Check Your Skills
- Chapter 22 Exercises

Chapter 23 Inference for Regression
- The regression parameters
- Testing the hypothesis of no linear relationship

  - Testing lack of correlation
  - Confidence intervals for the regression slope
  - Inference about prediction
  - Checking the conditions for inference
  - Chapter 23 Summary
  - This Chapter in Context
  - Check Your Skills
  - Chapter 23 Exercises
- Chapter 24 One-Way Analysis of Variance: Comparing Several Means
  - Comparing several means
  - The F statistic
  - The analysis of variance F test
  - Conditions for ANOVA
  - Details of ANOVA calculations
  - Chapter 24 Summary
  - This Chapter in Context
  - Check Your Skills
  - Chapter 24 Exercises
- Chapter 25 Statistical Inference: Part III Review
  - Essential concepts from Parts I and II
  - Part III summary
  - Comprehensive review exercises: inference selection
  - Comprehensive review exercises: analysis and conclusion
  - Large data set exercises
  - Online data sources
  - EESEE case studies

Back Matter

1. Notes and Data Sources
2. Tables
   TABLE A: Random Digits

  - TABLE B: Standard Normal Probabilities
  - TABLE C: *t* Distribution Critical Values
  - TABLE D: Chi-square Distribution Critical Values
  - TABLE E: Critical Values of the Correlation *r*
  - TABLE F: *F* Distribution Critical Values
  - TABLE G: Critical values m* for Tukey pairwise multiple comparisons with 95% confidence level, *k* comparisons, and *N* – *k* degrees of freedom (df)
3. Answers to Selected Exercises
4. Some Studies Recurring Across Chapters
5. Index
    1. A
    2. B
    3. C
    4. D
    5. E
    6. F
    7. G
    8. H
    9. I
    10. L
    11. M
    12. N
    13. O
    14. P
    15. Q
    16. R
    17. S
    18. T
    19. U
    20. V

    21. W
    22. X
    23. Z
6. Symbols

Part IV Optional Companion Chapters **(available on the *PSLS 4e* Companion Website)**

1. Chapter 26 More about Analysis of Variance: Follow-up Tests and Two-Way ANOVA
    - Beyond one-way ANOVA
    - Follow-up analysis: Tukey pairwise multiple comparisons
    - Follow-up analysis: contrasts
    - Two-way ANOVA: conditions, main effects, and interaction
    - Inference for two-way ANOVA
    - Some details of two-way ANOVA
    - Chapter 26 Summary
    - Statistics in Summary
    - This Chapter in Context
    - Check Your Skills
    - Chapter 26 Exercises
2. Chapter 27 Nonparametric Tests
    - Comparing two samples: the Wilcoxon rank sum test
    - Matched pairs: the Wilcoxon signed rank test
    - Comparing several samples: the Kruskal-Wallis test
    - Chapter 27 Summary
    - Statistics in Summary
    - This Chapter in Context
    - Check Your Skills
    - Chapter 27 Exercises
3. Chapter 28 Multiple and Logistic Regression
    - Parallel regression lines
    - Estimating parameters

    - Using technology
    - Conditions for inference
    - Inference for multiple regression
    - Interaction
    - A case study for multiple regression
    - Logistic regression
    - Inference for logistic regression
    - Chapter 28 Summary
    - Statistics in Summary
    - This Chapter in Context
    - Check Your Skills
    - Chapter 28 Exercises
- Back Cover

Landmarks

- Cover
- Begin Reading

List of Illustrations

- Figure 1.1 Example of a botanical database displayed in a spreadsheet. Each column includes data about a different variable. Each row represents data for one plant specimen.
- Figure 1.2 You can use either a pie chart or a bar graph to display the distribution of a categorical variable. Here is a pie chart of the reported cases of infectious diseases in California for 2014.
- Figure 1.3 The data on the reported cases of infectious diseases in California for 2014 are displayed in a bar graph format. (a) The bars are sorted according to their height (by order of importance). (b) The bars in the graph are grouped by disease type.
- Figure 1.4 Bar graph showing the percent of American adults in each of

four age groups who had serious suicidal thoughts in 2014. (a) The graph displays only those individuals in each group who had serious suicidal thoughts. (b) The graph displays both individuals who did and did not have serious suicidal thoughts in each age group, so that the stacked bars make up 100% of each age group.

Figure 1.5 Bar graph showing the percent of American adults who had serious suicidal thoughts in 2014, for each of four ethnic groups and for each of four age groups.

Figure 1.6 Histograms of the body length in feet for 44 great white sharks. (a) This histogram has 7 classes. (b) This histogram has 10 classes and shows more detail.

Figure 1.7 Histogram of 2014 rates of pertussis (in number of cases per 100,000 persons) in the 59 California counties. Classes include the lower-bound value but exclude the upper-bound value.

Figure 1.8 Histogram of patient age in years for 241,931 cases of Lyme disease reported in the United States between 1992 and 2006. Notice the two separate peaks around 10 years and 45 years.

Figure 1.9 Histogram of age (in years) at onset of anorexia nervosa for 691 Canadian girls diagnosed with the disorder. The first class includes 11-year-old girls but excludes 12-year-olds.

Figure 1.10 Dotplot of body length (in feet) for 44 great white sharks. (a) This dotplot is scaled horizontally, with similar values shown as stacked dots. (b) This dotplot is scaled vertically, with similar values distinguished by adding jitter between the dots.

Figure 1.11 Time plot of Shark River water salinity in Everglades National Park over a seven-day period in 2009. The daily cycles reflect the influence of the tides from the Gulf of Mexico, into which the Shark River discharges.

Figure 1.12 Histogram (a) and time plot (b) of the annual global temperature anomaly over the 133 years from 1880 to 2012. Data are from Table 1.1. Anomalies are relative to the 1951–1980 reference period.

Figure 1.13 Time plot of monthly atmospheric CO2 levels recorded at Mauna Loa Observatory, from March 1958 to January 2013.

Figure 1.14 Percent of adults who are overweight or obese in the world,

by country income level.
Figure 1.15 Histogram of the takeoff angles (measured in degrees) of 51 flea jumps.
Figure 1.16 Dotplot of the percent of population younger than 18 years of age for the 50 states and D.C.
Figure 1.17 Histograms of scores on a quiz of past public events (a) for 115 individuals claiming to have Highly Superior Autobiographical Memory (HSAM) and (b) for 30 control individuals.
Figure 1.18 Percent of the adult population who is obese, plotted for each state on a map of the United States
Figure 1.19 Time plot of the percent of respondents in numerous Gallup polls who say that they exercised at least 30 minutes on three or more days in the previous week.
Figure 2.1 Output from a graphing calculator, a spreadsheet program, and four software packages describing the Aleppo pine needle length data.
Figure 2.2 Dotplot of the number of flood events recorded over the 2009-2013 period for each of 52 coastal communities.
Figure 2.3 Metabolic rates for 7 men, with their mean (*) and the deviations of two observations from the mean indicated.
Figure 2.4 Boxplots comparing the lengths of 15 Aleppo pine needles and 18 Torrey pine needles.
Figure 2.5 The same data as in Figure 2.4 now displayed as the mean (columns) plus and minus one standard deviation (error bars) overlaid on top of a dotplot of the raw data for both species.
Figure 2.6 Dotplot of the sizes of 11 acorns.
Figure 2.7 Histogram and modified boxplot (using the same horizontal scale) of the number of flood events recorded over the 2009-2013 period for each of 52 coastal communities.
Figure 2.8 Dotplots comparing the distributions of scores on a test of verifiable autobiographical memory.
Figure 2.9 Boxplots of the distributions of number of aggressions received per half hour, for female macaque monkeys of differing social status.

Figure 3.1 Scatterplot of the number of manatee deaths due to powerboat collisions in Florida each year against the number of powerboats registered (in thousands) that same year. The dotted lines intersect at the point (755, 54), the data for year 1997.

Figure 3.2 Scatterplot of fat gain after 8 weeks of overeating plotted against increase in nonexercise activity over the same period.

Figure 3.3 Scatterplot of the relationship between starvation and fledging in Magellanic penguin chicks between 1983 and 2010. Data for two years with extremely unusual weather are shown as open symbols.

Figure 3.4 Cost of reproduction: scatterplot of thorax length (in millimeters) and longevity (in days) for sexually active (triangles) or inactive (circles) male fruit flies.

Figure 3.5 How correlation measures the strength of a linear relationship, for Example 3.8. Patterns closer to a straight line have correlations closer to 1 or −1.

Figure 3.6 Scatterplot of Gesell Adaptive Score versus the age at first word for 21 children.

Figure 3.7 Scatterplot of energy expenditure against running speed, using different colors and symbols to represent various running inclines.

Figure 3.8 Scatterplot of the percent of animals reacting to a noxious luminous stimulus consisting of varying numbers of consecutive light pulses. Circles represent wild-type C. elegans, and squares represent mutant C. elegans.

Figure 4.1 Fat gain after 8 weeks of overeating, plotted against increase in nonexercise activity over the same period.

Figure 4.2 Least-squares regression for the nonexercise activity data: output from a graphing calculator, three statistical programs, and a spreadsheet program.

Figure 4.3 Scatterplot of activity in a region of the brain that responds to pain versus score on a test of empathy, for Example 4.45. Brain activity is measured as the subject watches her partner experience pain. The line is the least-squares regression line. Subject 16 is an outlier in the x direction.

Figure 4.4 Variations on Figure 4.3.(a) for Example 4.5. (a) Subject 16 is an outlier in the x direction. It is not influential for least-squares regression, because removing it moves the regression line (dashed line) only a little. (b) An outlier in the x direction pulls the least-squares line to itself when there are no other observations with similar values of x to hold the line in place. Altering Subject 16’s data to move its value down pulls the regression line down (dashed line).

Figure 4.5 (a) Scatterplot of body weight and brain weight for 96 species of mammals in the original numerical values of kilograms and grams, for Example 4.6. (b) Scatterplot of the data in (a) using the logarithm of brain weight and the logarithm of body weight.

Figure 4.6 JMP regression output for the cancer study of Exercise 4.9.

Figure 4.7 (a) Scatterplot of sympathetic nerve activity and an indicator of arterial stiffness in 44 healthy young adults, for Example 4.8. Notice the lack of any obvious association. (b)–(c) Scatterplot of the data in (a) after separating the data for men (b) and for women (c). Notice the opposite trends for men (positive linear association) and for women (negative linear association).

Figure 4.8 Outside temperature and frequency of mating calls for 16 tree frogs, for Exercise 4.15.

Figure 4.9 Minitab regression output for a study of the effect of temperature on toucan beak heat loss, for Exercise 4.30.

Figure 4.10 CrunchIt! regression output for a study of the effects of social rejection on brain activity, for Exercise 4.31.

Figure 5.1 A bar graph of the distribution of shoes worn by the type of structural support they offer. This is one of the marginal distributions for Table 5.1.

Figure 5.2 (a) Bar graph comparing the percents of shoe types (offering good, average, and poor structural support) among men and among women in the study. Men and women wore all three types of shoes, but the percent of shoes offering poor structural support was higher among the women. (b) The same percents are displayed in a “stacked” bar graph, which helps visualize the fact that the conditional distribution within each gender adds up to 100%.

Figure 5.3 Minitab and R outputs of the two-way table of shoe support types for men and women with each entry as a percent of its column total. The percents in each column give the conditional distribution of shoe support type for one gender. The “All” column in the Minitab output also gives the marginal distribution of shoe support type for the whole study (both genders combined).

Figure 5.4 Prevalence of alcohol dependence among U.S. adult binge drinkers, by binge drinking frequency during the past 30 days.

Figure 6.1 Confounding: We cannot distinguish the effects of what people drink from the effects of their overall lifestyle, diet, and socioeconomics.

Figure 6.2 Analogy for taking samples. Four samples are shown. Imagine guessing the location of the center of the population from the center of the sample distribution.

Figure 6.3 The samples from Figure 6.2 are now superimposed on archery targets representing the target population. Sample (d) in this analogy is a biased sample.

Figure 7.1 The treatments in the experimental design for studying flowering. Combinations of values of the two factors form 12 treatments.

Figure 7.2 Confounding. We cannot distinguish the effect of the treatment from the effects of lurking variables in the uncontrolled experiment about gastric freezing.

Figure 7.3 Outline of a randomized comparative experiment to compare the effect of acupuncture, sham acupuncture, and placebo on headaches.

Figure 7.4 Outline of a block design comparing two asthma care options for asthmatic children who are or are not exposed to smoke at home.

Figure 7.5 Outline of a block design. The blocks are land areas subdivided into six planting plots. The treatments are six combinations of tillage type and pesticide application.

Figure 7.6 Data collection for an experiment with two conditions, A and B, under a completely randomized design (left) versus a matched pairs design (right).

Figure 8.1

Figure 8.2

Figure 8.3

Figure 8.4

Figure 8.5

Figure 9.1 The proportion of tosses of a coin that give a head changes as we make more tosses. Eventually, however, the proportion approaches 0.5, the probability of a head. This figure shows the results of two trials of 5000 consecutive tosses each.

Figure 9.2 Histogram of the heights in inches of women age 40 to 49 in the United States. The smooth curve shows the overall shape of the distribution.

Figure 9.3 (a) The proportion of women 62 inches tall or less in the sample is 0.319. (b) The proportion of women 62 inches tall or less calculated from the density curve is 0.316. The density curve is a good approximation to the distribution of the data.

Figure 9.4 Probability as area under a density curve. This uniform density curve spreads probability evenly between 0 and 1.

Figure 9.5 Distribution of age of onset of anorexia nervosa, pictured by both a histogram and a density curve.

Figure 10.1 Venn diagram showing disjoint events A and B. Their joint probability, P(A and B), equals zero.

Figure 10.2 Venn diagram showing events A and B that are not disjoint. The event {A and B} consists of outcomes common to A and B.

Figure 10.3 The general addition rule: P(A or B)=P(A)+P(B)-P(A and B) for any two events A and B.

Figure 10.4 Venn diagrams and probabilities for hearing impairment and eye color in Dalmatian dogs. (a) This Venn diagram displays the simple probabilities of I and B. (b) This Venn diagram displays the probabilities for all possible combinations of I and B. Notice that, in both diagrams, the probability values shown match their corresponding labels.

Figure 10.5 Tree diagram of skin cancer. The two stages are the location on the body of the cancer and the gender of the cancer patient.

Figure 10.6 Tree diagram of the features of children of Caucasian descent in Germany. The three stages are hair color, eye color, and whether or not the child has freckles.

Figure 10.7 All possible outcomes of a diagnostic test, represented with (a) a two-way table and (b) a tree diagram.

Figure 10.8 (a) Tree diagram for mammograms given to women in their 60s. (b) Corresponding two-way table.

Figure 10.9 Kaiser Permanente’s outcome diagram of 100 PSA tests. Each dot represents the result of one PSA test.

Figure 11.1 Two Normal curves with a different mean μ and a different standard deviation σ.

Figure 11.2 The 68–95–99.7 rule for Normal distributions.

Figure 11.3 The 68–95–99.7 rule applied to the distribution of heights of young women age 18 to 24 years, with $\mu=64.5$ inches and $\sigma=2.5$ inches.

Figure 11.4 Minitab graph displaying the area to the left of −2.5 under the N(-2,1) density curve. This area, 0.3085, represents the probability that a randomly selected woman in her 70s would have osteoporosis.

Figure 11.5 Software output for the cumulative probability up to −2.5 for the N(-2,1) density curve.

Figure 11.6 JMP output for the area between −2.5 and −1 under the N(-2,1) distribution. The shaded area is 0.5328 and represents the proportion of older women diagnosed with osteopenia.

Figure 11.7 Obtaining the probability of osteopenia in older women (area between −2.5 and −1) using cumulative probabilities.

Figure 11.8 Locating the third quartile of a Normal curve. The third quartile x is the 75th percentile.

Figure 11.9 Software output for the third quartile of the N(45,4) density curve used to model the distribution of hatching weights.

Figure 11.10 Finding the 10% longest pregnancies. The value x we need to find is the 90th percentile of pregnancy lengths.

Figure 11.11 The area under a standard Normal curve to the left of the point $z=1.47$ is 0.9292. Table B gives cumulative areas under the standard Normal curve.

Figure 11.12 Distribution of shark lengths. (a) Vertical dotplot with jitter added for visibility, from Figure 1.10b. A low outlier and a high outlier are evident. (b) Normal quantile plot of shark lengths, which shows an approximately Normal distribution except for a low outlier and a high outlier.

Figure 11.13 (a) Histogram and (b) Normal quantile plot of guinea pig survival times. This distribution is skewed to the right.

Figure 11.14 (a) Histogram with superimposed Normal curve and (b) Normal quantile plot of rainwater pH values. This distribution is roughly Normal.

Figure 11.15 Normal quantile plot of the heart rates of 200 male runners.

Figure 11.16 Normal quantile plots of (a) seed count for 19 tree species and (b) logarithm of seed count for these tree species.

Figure 11.17 Normal quantile plot of the volumes of 28 acorns from the Atlantic region.

Figure 11.18 Normal quantile plot of age at onset of anorexia nervosa in 691 teenage girls.

Figure 11.19 Normal quantile plot of the thorax lengths of 49 male fruit flies.

Figure 12.1 The binomial probability P(X≤1): output from a graphing calculator and three computer programs.

Figure 12.2 Probability histogram for the binomial distribution with n=10 and p=0.1.

Figure 12.3 Probability distribution for the binomial model with n = 2000 and p = 0.62, displayed graphically. The height of each bar represents the probability for X when it takes a value on the horizontal axis. The yellow bars represent probabilities for values of X≥1250. Notice how the shape of this binomial probability distribution closely resembles a Normal curve.

Figure 12.4 Graphical display of a simulated field containing dandelions (circles). Dandelion location on the rectangular field follows a Poisson distribution, with mean 2.1 dandelions per quadrant.

Figure 12.5 The Poisson probability P(X≤1): output from three computer programs and a graphing calculator.

Figure 12.6 CrunchIt! output for the Poisson probability P(X≥4).

Figure 13.1 Simulating random samples from the N(100, 15) distribution of IQ scores. (a) The population distribution N(100, 15) is

displayed graphically. (b) Each of the first 10 rows represents one simulated random sample of 25 individuals from the N(100, 15) population. The last row shows the means $\bar{x}$ of these 10 simulated random samples.

Figure 13.2 The idea of a sampling distribution: Take many SRSs from the same population, compute their sample means $\bar{x}$, and display the distribution of $\bar{x}$. The histogram shows the result of 500 simulated SRSs from the N(100, 15) population distribution of IQ scores.

Figure 13.3 Population distribution N(100, 15) at the top. Each row underneath it shows the means $\bar{x}$ of 10 simulated SRSs in each of four situations varying by the chosen sample size n.

Figure 13.4 The distribution of single observations (the population) compared with the distribution of the means $\bar{x}$ of 25 observations (the sampling distribution). Because the population has a Normal distribution, the sampling distribution of $\bar{x}$ is also Normal, but it is narrower. Averages are less variable than individual observations.

Figure 13.5 (left) The central limit theorem in action: The distribution of sample means $\bar{x}$ from a strongly non-Normal population is more Normal for larger sample sizes. (a) The distribution of single observations (the population distribution). (b) The sampling distribution of $\bar{x}$ for samples of 2 observations. (c) The sampling distribution of $\bar{x}$ for samples of 10 observations. (d) The sampling distribution of $\bar{x}$ for samples of 25 observations.

Figure 13.6 (right) The sampling distribution of $\bar{x}$ is always Normal when the population is Normal. (a) The distribution of single observations (the population distribution IQ scores for a particular age group). (b) The sampling distribution of $\bar{x}$ for samples of 2 observations. (c) The sampling distribution of $\bar{x}$ for samples of 10 observations. (d) The sampling distribution of $\bar{x}$ for samples of 25 observations.

Figure 13.7 Histograms of three data sets for Example 13.6. (a) The angle of big-toe deformations (hallux abducto valgus [HAV]) in 38 patients. (b) The number of daily fruit servings for 74 adolescent girls. (c) The lengths of 56 perch from a Swedish lake.

Figure 13.8 The idea of a sampling distribution: Take many SRSs from the same population, compute their sample proportions p, and display

the distribution of p. The histogram shows the result of 500 simulated SRSs from the population of blood types, with proportion 0.45 having type O.

Figure 13.9 Select a large SRS from a population in which a proportion p are successes. The sampling distribution of the proportion p of successes in the sample is approximately Normal. The mean is p, and the standard deviation is (p(1-p)/n).

Figure 14.1 The sampling distribution of the mean height x¯ of an SRS of 217 eight-year-old boys. In 95% of all samples, x¯ lies within ± 1.4 of the unknown population mean μ. To say that x¯± 1.4 is a 95% confidence interval for the population mean μ is to say that, in repeated samples, 95% of those intervals capture μ.

Figure 14.2 Twenty-five SRSs from the same population gave these 95% confidence intervals. In the long run, 95% of all samples give an interval that contains the population mean μ.

Figure 14.3 The critical value z* is the number that catches central probability C under a standard Normal curve between −z* and z*.

Figure 14.4 Dotplot of the IQ scores of 31 seventh-grade girls.

Figure 14.5 If people age 75 to 79 do not differ from the whole adult population in their blood levels of inorganic phosphorus, the mean inorganic phosphorus level x¯ for a random sample of 12 individuals age 75 to 79 will have this sampling distribution. The actual result for this group is x¯= 1.128. That point is so far out on the Normal curve that it is good evidence that people age 75 to 79 do have inorganic phosphorus levels lower than 1.2 mmol/l.

Figure 14.6 Output from the P-Value of a Test of Significance applet for Example 14.8. The P-value for the one-sided test, 0.0063, is the area under the Normal curve to the left of x¯=1.128.

Figure 14.7 Output from the P-Value of a Test of Significance applet for Example 14.9. The P-value for the two-sided test, 0.2295, is twice the value of the area under the Normal curve to the right of x¯=326.9.

Figure 14.8 Software output (JMP) for the two-sided test in Example 14.9. The observed value of the test statistics is z=- 6.65, and the two-sided P-value is roughly zero.

Figure 15.1 The lengths of three confidence intervals for the IQ study.

All three are centered at the estimate x-=105.84. When the data and the sample size remain the same, higher confidence requires a larger margin of error.

Figure 15.2 Output from the Statistical Power applet for the cola study with $\sigma=1$, $n=10$, and significance level 0.05. The power of the test $H0:\mu=0$ against the specific alternative $Ha:\mu=0.8$ is 0.812.

Figure 15.3 Output from Minitab for the cola study with $\sigma=1$ and significance level 0.05. The sample size required to reach a target power of 0.9 for the one-sided z test aiming for various effect sizes, here labeled “difference” (the difference between μ under H0 and Ha).

Figure 15.4 The two types of error in testing hypotheses and their probability.

Figure 16.1 Statistics in summary: The Idea of a confidence interval.

Figure 16.2 Statistics in summary: The idea of a hypothesis test.

Figure 16.3 Density curve for the location of accidents along a 3-mile bike path.

Figure 17.1 Density curves for the t distributions with 5, 10, and 20 degrees of freedom (solid lines) and for the standard Normal distribution (dotted line). All are symmetric with center 0. The t distributions are somewhat more spread out.

Figure 17.2 Dotplot of plasma testosterone levels in obese adolescent males.

Figure 17.3 The t confidence interval for the testosterone data: output from a graphing calculator, three statistical programs, and a spreadsheet program.

Figure 17.4 Software output (JMP) for the forewing weight of monarch butterflies.

Figure 17.5 Dotplot of pulse wave velocity (PWV) in children with progeria.

Figure 17.6 The t test for the study of progeria: output from a graphing calculator and four statistical programs.

Figure 17.7 Software output (Minitab) for two wearable fitness devices.

Figure 17.8 Dotplot (with jitter) of PWV differences for 18 children with progeria.

Figure 17.9 Software output (R) for the essential tremor matched pairs experiment.

Figure 17.10 Can we use t procedures for these data? (a) Volumes of 11 oak tree acorns. No, there is a clear outlier and the sample size is small. (b) Lengths of 15 Aleppo pine needles. Yes, the sample is large enough to overcome the mild skewness. (c) Weights of roughly 500 brown treesnakes. Yes, the very large sample size is enough to overcome the pronounced skewness and the extreme outlier, although the impact of the outlier should be given careful consideration. (d) Widths of roughly 800 bee cocoons. No, because the distribution is clearly bimodal, so it cannot be reasonably described by just one mean.

Figure 18.1 Dotplot of the time spent walking or standing.

Figure 18.2 The two-sample t procedures applied to the data on activity and obesity: output from a graphing calculator, three statistical programs, and a spreadsheet program.

Figure 18.3 The P-value as the upper-tail area, for Example 18.3. We rounded down the degrees of freedom to 15.

Figure 18.4 Boxplots of hatching weights of commercial and transgenic chickens.

Figure 18.5 Software output (Minitab) for chicken hatching weights.

Figure 18.6 The P-value for Example 18.5. Because the alternative is two-sided, the P-value is double the area to the right of t = 0.19. We rounded the degrees of freedom to 95.

Figure 18.7 Software output (JMP) for C4A expression levels in patients with and without schizophrenia.

Figure 19.1 With probability C, the random variable p^ lies within ±z*p(1-p)/n of the unknown population proportion p. In other words, in these samples p lies within ±z*p(1-p)/n of p^.

Figure 19.2 Output from a graphing calculator and three statistical software programs for the 95% confidence interval for the proportion of women ages 14 to 59 who test positive for HPV.

Figure 19.3 When H0:p=p0 is true, the sampling distribution of p^ is approximately Normal with mean p0 and standard deviation p0(1-p0)/n.

Figure 19.4 Output from a graphing calculator, two software packages, and an applet for the hypothesis test about the righting behavior of aphids.

Figure 20.1 Select independent SRSs from two populations having proportions of successes p1 and p2. The proportions of successes in the two samples are p1 and p2 When the samples are large, the sampling distribution of the difference p1-p2 is appropriately Normal.

Figure 20.2 Output from a graphing calculator and three statistical programs for a study comparing the success rates of abrupt and gradual smoking cessation.

Figure 20.3 Output from a graphing calculator and three statistical programs for a study of the efficacy of a new HIV vaccine.

Figure 20.4 The P-value for the one-sided test for a study of the efficacy of a new HIV vaccine.

Figure 20.5 JMP output for a study of the association between HPV and oropharyngeal cancer.

Figure 21.1 Density curves for the chi-square distributions with 1, 4, and 8 degrees of freedom. Chi-square distributions take only positive values and are right-skewed.

Figure 21.2 Output from JMP, Minitab, R, and Excel for a study of amaranth seed color.

Figure 21.3 Minitab output for a study of birth days.

Figure 21.4 Distribution of pushing arm in a moving octopus. The numbers 1 through 8 refer to the clockwise location of the arm on the body, starting with the front right location (arm 1).

Figure 21.5 Minitab output for Exercise 21.29.

Figure 21.6 Excel calculations for Exercise 21.36.

Figure 22.1 Bar graph comparing hormone therapy (ERT) history for postmenopausal women with (cases) or without (controls) a first diagnosis of thromboembolism.

Figure 22.2 Output from the TI-83 graphing calculator, Minitab, JMP, and R Commander for the chi-square test in the thromboembolism study.

Figure 22.3 Minitab chi-square output for the two-way table of e-cigarettes and smoking cessation.

Figure 22.4 Minitab output for the two-way table in the online harassment study.

Figure 22.5 Minitab output for the two-way table of HPV status by age group among American women.

Figure 22.6 SPSS output for the two-way table in the study of hearing impairment among Dalmatian dogs.

Figure 22.7 Minitab output for the study of acne treatments.

Figure 22.8 Minitab output for the two-way table in the forest study.

Figure 22.9 Minitab output for the experimental results of the study of stroke prevention.

Figure 22.10 JMP output for the melanoma epidemiological data.

Figure 23.1 Scatterplot of blood-brain barrier (BBB) permeability in the hippocampus of 24 healthy adults as a function of their age.

Figure 23.2 Regression of BBB permeability on age: output from a graphing calculator, three statistical programs, and a spreadsheet program.

Figure 23.3 The nature of regression data when the conditions for inference are met. The line is the population regression line, which shows how the mean response μy changes as the explanatory variable x changes. For any fixed value of x, the observed response y varies according to a Normal distribution having mean μy and standard deviation σ.

Figure 23.4 Minitab output for the frog mating-call data.

Figure 23.5 Scatterplot of the reading-skill scores and nonverbal IQ scores of 22 dyslexic children.

Figure 23.6 Scatterplot of hand lengths and body heights for 21 men.

Figure 23.7 Scatterplot of college students' blood alcohol content (BAC) against the number of cans of beer consumed, with the least-squares regression line.

Figure 23.8 Minitab output for the blood alcohol content data, with prediction for x*=5.

Figure 23.9 Residual plot (a) and histogram of residuals (b) for the regression of BBB permeability on age.

Figure 23.10 Scatterplot of blood-brain barrier permeability in the hippocampus of 24 healthy adults as a function of their age. The two regression lines are for the data with (solid line) and without (dashed line) the unusual observation (open circle).

Figure 23.11 Minitab output for the nonexercise activity (NEA) change and fat gain data, with prediction for x*=500.

Figure 23.12 Scatterplot of the log of offspring nest size against the log of parent nest size.

Figure 23.13 Partial Minitab output for predicting number of new birds in a sparrowhawk colony from percent of birds returning.

Figure 24.1 Dotplots comparing the body weights in grams of rats randomly assigned to three diet groups.

Figure 24.2 Bar graph comparing the mean rest time of fruit flies given different amounts of caffeine.

Figure 24.3 Density curves for three F distributions: F(2, 30), F(5, 30), and F(5, 120). F distributions take only positive values and are right-skewed.

Figure 24.4 ANOVA for the rat weight data: output from a graphing calculator, three statistical programs, and a spreadsheet program.

Figure 24.5 Excel ANOVA output for the number of trees in forest plots in Borneo.

Figure 24.6 Dotplots comparing the meconium cotinine levels of infants born to mothers with different smoking statuses.

Figure 24.7 Minitab ANOVA output for comparing the total plant biomass of grassland plots under different water conditions.

Figure 25.1

Figure 25.2

Figure 25.3 Summary statistics provided in a study of autism spectrum disorder (ASD).

Figure 26.1 Minitab ANOVA output for the rat weight data of Example 24.3.

Figure 26.2 Additional Minitab ANOVA output showing Tukey pairwise multiple comparisons for the rat weight data, for Example 26.2.

Figure 26.3 Minitab ANOVA output for the study on attracting cereal leaf beetles, for Example 26.6.

Figure 26.4 Plot of the means for the made-up data from Example 26.8. In addition to the means for each of the six conditions, the plot displays the average for each of the low, medium, and high dosage groups as a dotted line. The parallel lines show that there is no interaction between the two factors.

Figure 26.5 Plot of the means for the made-up data from Example 26.9, along with the averages for each of the low, medium, and high dosage groups. The lines are not parallel, so there is an interaction between delivery method and dosage.

Figure 26.6 Plots of two sets of means for a made-up study comparing the performance of normal and hyperactive children under two conditions, for Example 26.10. The lines in (a) are not parallel, indicating an interaction between the two factors. In addition to the interaction effect, the large gap between the two lines in (b) indicates a very strong main effect of whether a child is hyperactive or not.

Figure 26.7 Plots of two sets (a and b) of possible means (solid lines) for a study in which the two factors are soybean variety and type of herbicide. The plots also show the average for each herbicide (dotted line).

Figure 26.8 Side-by-side dotplots comparing the percent of body fat of wild-type and mutant female fruit flies for four different amounts of yeast in the diet, for Example 26.11. Genotype 0 = mutant; Genotype 1 = wild-type.

Figure 26.9 Plots of the group means from the fruit fly study, from Minitab. The two plots use the same eight means. They differ only in the choice of which variable to mark on the horizontal axis.

Figure 26.10 Partial two-way ANOVA output from Minitab for the fruit fly study data in Table 26.1.

Figure 26.11 Plots from Minitab of the group means for the study of phosphorus levels in tomatoes, for Example 26.12.

Figure 26.12 Plots from Minitab of the mean weights of 21-day-old chicks fed 9 different diets, for Example 26.13.

Figure 26.13 Two-way ANOVA output from Minitab for the study of the effect of diet on chick growth, for Example 26.13.

Figure 26.14 Compare the sums of squares in these one-way and two-way ANOVA outputs for the fruit fly data from Table 26.1.

Figure 26.15 Two-way ANOVA table for Exercises 26.20 to 26.23.

Figure 26.16 Plots from Minitab of the mean activity of a primary visual cortex neuron under six conditions of disparity and three visual distances, for Exercise 26.34.

Figure 27.1 Comparison of tests based on Normal distributions with rank tests for similar settings.

Figure 27.2 Dotplot of corn yields from plots with no weeds and with 3 weeds per meter of row.

Figure 27.3 Output from R for the data in Example 27.1. The output compares the three methods for computing the P-value of the Wilcoxon rank sum test and the P-value obtained with a two-sample t test.

Figure 27.4 Output from Minitab for the data of Example 27.5.

Figure 27.5 Output from Minitab, CrunchIt!, SPSS, and R for the neural activity data of Example 27.6. CrunchIt! uses the exact method for calculating P, whereas Minitab and SPSS use a Normal approximation (one with and the other without a continuity correction). R gives the option of selecting the computation method (all three are included for comparison).

Figure 27.6 Dotplot of the differences in serum creatinine level before and after treatment, for Example 27.9.

Figure 27.7 Output from Minitab for the serum creatinine data of Example 27.9. Because there are ties, a Normal approximation must be used for the Wilcoxon signed rank test. The results of a paired t test are provided for comparison.

Figure 27.8 Minitab output for the corn yield data of Example 27.10. For comparison, both the Kruskal-Wallis test and the one-way ANOVA are shown.

Figure 28.1 A scatterplot of longevity of male fruit flies of different adult sizes, with two separate regression lines, for Example 28.1.

Figure 28.2 Multiple regression model with two parallel straight lines, for Example 28.2.

Figure 28.3 Output from Excel, Minitab, SPSS, and R for the model with parallel regression lines in Example 28.3.

Figure 28.4 Residual plots to check the conditions for inference for the model with parallel regression lines in Example 28.3.

Figure 28.5 Scatterplot and residual plot to check the conditions for inference in Exercise 28.5.

Figure 28.6 Residual plot and histogram of the residuals to check the conditions for inference in Exercise 28.6.

Figure 28.7 Predicted values, confidence limits for mean longevity, and prediction limits for future longevity for two fruit fly observations, for Example 28.11.

Figure 28.8 Scatterplot of FEV versus age for male and female children, for Example 28.12.

Figure 28.9 Output from Minitab for the model with two regression lines in Example 28.13.

Figure 28.10 Some possible surfaces for multiple regression models, for Example 28.14. (a) Plane μy=x1+x2.$\mu_y = x_1 + x_2$. (b) Surface μy=2x1+2x2+10x1x2.$\mu_y = 2x_1 + 2x_2 + 10x_1x_2$. (c) Surface μy=2000-20x12-2x1-3x22+5x2+10x1x2.$\mu_y = 2000 - 20x_1^2 - 2x_1 - 3x_2^2 + 5x_2 + 10x_1x_2$.

Figure 28.11 Scatterplots of energy expenditure against (a) speed and against (b) incline for Example 28.15.

Figure 28.12 Scatterplots of energy expenditure against speed and incline using (a) a 3D-surface plot or a (b) 2D plot with regression lines of different colors for various inclines.

Figure 28.13 Output from Minitab for the model with two variables and their interaction term in Example 28.17.

Figure 28.14 (a) Linear relationship between the logit L and an explanatory variable x and (b) its equivalent when L is untransformed to obtain p, the probability of a given outcome. The relationship between p and x has a sigmoidal, or "S," shape.

Figure 28.15 Dotplots of white cell counts (per mm^3 of cerebrospinal fluid) in patients with viral (abm = 0) or bacterial (abm = 1) meningitis, for Example 28.18. Note that each dot represents up to seven observations.

Figure 28.16 Minitab output for the simple logistic regression model in Example 28.18.

Figure 28.17 Minitab output for Exercise 28.42.
Figure 28.18 Minitab output for Exercises 28.49 to 28.51.

List of Tables

Table 1.1 Annual global temperature anomaly (degrees Celsius)
Table 1.2 Omega-3 fatty acids as a fraction of omega-6 fatty acids in food oils
Table 1.3 Percent of the population who is obese, for each state and the District of Columbia
Table 1.4 Number of unique cases of herbicide resistance in weeds worldwide, by year
Table 2.1 Time (minutes per day) active and lying down by lean and obese subjects
Table 3.1 Powerboat registrations (in thousands) and manatee deaths from powerboat collisions in Florida
Table 3.2 CO2 emissions (in ton per capita) and climate change concern index
Table 3.3 Actual volume and computer reconstruction of 30 oysters
Table 3.4 Annual average daily temperature and annual percent of α-acids in hops
Table 3.5 Count and weight of seeds produced by common tree species
Table 3.6 Lifetime risk and number of stem cell divisions for 31 cancer types
Table 3.7 Fish supply and wildlife decline in West Africa
Table 4.1 Four data sets for exploring correlation and regression
Table 4.2 Size and abundance of carnivores
Table 5.1 Study participants by shoe type and gender
Table 5.2 Shoe type by gender: men’s group
Table 5.3 Gender by shoe type: good support group
Table 8.1 Days from April 20 for the Tanana River tripod to fall
Table 8.2 Angle of deformity (degrees) for two types of foot deformity

Table 8.3 Neuron response (spikes per second) to tones and monkey calls

Table 14.1 Heights in inches of female undergraduate students at University of California at Irvine

Table 17.1 PWV (m/s) of 18 children with progeria, before and after treatment

Table 17.2 Glucose level (mg/dl) of 32 adolescents with type 1 diabetes under two treatment conditions, administered to each subject in random order

Table 17.3 Trials (out of 24) in which chimpanzees recruited a partner

Table 17.4 Absorption extent for two versions of a drug

Table 18.1 Time (minutes per day) spent in three different postures by lean and obese subjects

Table 18.2 Hatching weights (g) of transgenic and commercial chickens

Table 18.3 Weight change (lb) of obese children, by group

Table 18.4 Drinks per session claimed by female and male students

Table 18.5 Body weights (g) at beginning and end of ink toxicity study

Table 18.6 Longevity (days) of male *Drosophila*

Table 23.1 Age (in years) and BBB permeability (in 10^3) in the hippocampus of 24 healthy adults. The table also includes computations (predictions, residuals, and squared residuals) from a least-squares regression analysis of the data.

Table 23.2 Frog mating calls: predictions and residuals

Table 23.3 Actual volume and two computer reconstruction outputs for a sample of 30 oysters

Table 23.4 DNA and phytopigment concentrations (g/m^2) on the ocean floor

Table 23.5 Body mass index (BMI, in kg/m^2) and testosterone concentration (in nmol/l) of 50 adolescent males

Table 24.1 Body weights (in grams) of adult male rats fed one of three diets

Table 24.2 Data from a study of logging in Borneo

Table 24.3 Penetrability of soil at three compression levels

Table 24.4 Healing times (in days) for a deep surgical wound

Table 25.1 Measurements on 56 perch

Table 25.2 Healing rates μ when the skin voltage is either reduced (experimental limb) or left unchanged (control limb) in a sample of 14 anesthetized newts

Table 25.3 Flower lengths (millimeters) for three *Heliconia* varieties

Table 25.4 Ventilation rates (number of opercular beats per minute) of 90 goldfish kept at different temperatures (degrees Celsius)

Table 26.1 Percent lipid in female fruit flies by genotype and diet

Table 26.2 Percent of phosphorus in tomato plants

Table 26.3 Effect of electrical field on healing rate in newts

Table 26.4 Goldfish ventilation rates (number of opercular beats per minute)

Table 27.1 Average time (seconds) to complete a maze

Table 27.2 Mathematics skills scores before and after a subliminal message

Table 27.3 Electric fish species in the Amazon

Table 28.1 Thorax length (mm) and longevity (days) of reproducing and nonreproducing male fruit flies

Table 28.2 Sympathetic nerve activity (bursts/100 heartbeats) and arterial stiffness (mm Hg) of men and women

Table 28.3 Energy expenditure while running at various speeds (km/h) and treadmill inclines (%)

Table 28.4 Measurements on 56 perch

Table 28.5 Chemical composition and taste score of aged cheddar cheese

Table 28.6 Cybex treadmill display’s claimed Calories per hour by speed and incline for a 175-lb man

Table 28.7 Foot width and length of fourth-graders (in cm)

Table 28.8 World record running times for men and women

Table 28.9 Percent of phosphorus in tomato plants

Table 28.10 Reaction times (ms) in a computer game

Table 28.11 Measured and self-estimated reading ability data for 60

fifth-grade students randomly sampled from one elementary school
FC
i
ii
iii
iv
v
vi
vii
viii
ix
x
xi
xii
xiii
xiv
xv
xvi
xvii
xviii
xix
xx
xxi
xxii
xxiii
xxiv
xxv
xxvi
xxvii
xxviii
xxix

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31

32
33
34
35
36
37
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63

64

65

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

93

94

95

96

97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
212
213
215
216
217
218
219
220
221
222

223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315

316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346

347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
403
404
405
406
407
408
409

410
411
412
413
414
415
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441

442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566

567
568
569
570
571
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

721

722

723

724

725

726

727

729

26-1

26-2

26-3

26-4

26-5

26-6

26-7

26-8

26-9

26-10

26-11

26-12

26-13

26-14

26-15

26-16

26-17

26-18

26-19

26-20

26-21

26-22

26-23

26-24

26-25

26-26

26-27

26-28
26-29
26-30
26-31
26-32
26-33
26-34
26-35
26-36
26-37
26-38
26-39
26-40
27-1
27-2
27-3
27-4
27-5
27-6
27-7
27-8
27-9
27-10
27-11
27-12
27-13
27-14
27-15
27-16
27-17
27-18

27-19
27-20
27-21
27-22
27-23
27-24
27-25
27-26
27-27
27-28
27-29
27-30
27-31
27-32
27-33
27-34
27-35
27-36
28-1
28-2
28-3
28-4
28-5
28-6
28-7
28-8
28-9
28-10
28-11
28-12
28-13

28-14

28-15

28-16

28-17

28-18

28-19

28-20

28-21

28-22

28-23

28-24

28-25

28-26

28-27

28-28

28-29

28-30

28-31

28-32

28-33

28-34

28-35

28-36

28-37

28-38

28-39

28-40

28-41

28-42

28-43

28-44

28-45
28-46
28-47
28-48
28-49
28-50
28-51
28-52
28-53
28-54
28-55
28-56
28-57
28-58
28-59
BC

目录

Title Page 2

Copyright Page 3

Brief Contents 5

Contents 8

To the Instructor: About this Book 20

Acknowledgments 35

Media and Supplements 39

To the Student: Statistical Thinking 43

About the Authors 53

Part I: Collecting and Exploring Data 55

- Chapter 1 Picturing Distributions with Graphs 57
  - Individuals and variables 59
  - Identifying categorical and quantitative variables 61
  - Categorical variables: pie charts and bar graphs 67
  - Quantitative variables: histograms 77
  - Interpreting histograms 82
  - Quantitative variables: dotplots 89
  - Time plots 93
  - Discussion: (Mis)adventures in data entry 99
  - Chapter 1 Summary 105
  - This Chapter in Context 107
  - Check Your Skills 108
  - Chapter 1 Exercises 112
- Chapter 2 Describing Quantitative Distributions with Numbers 127
  - Measures of center: median, mean 129
  - Measures of spread: percentiles, standard deviation 139
  - Graphical displays of numerical summaries 148
  - Spotting suspected outliers 153
  - Discussion: Dealing with outliers 157
  - Organizing a statistical problem 163
  - Chapter 2 Summary 167

This Chapter in Context 169
Check Your Skills 170
Chapter 2 Exercises 172
Chapter 3 Scatterplots and Correlation 181
Explanatory and response variables 183
Relationship between two quantitative variables: scatterplots 186
Adding categorical variables to scatterplots 196
Measuring linear association: correlation 201
Chapter 3 Summary 210
This Chapter in Context 212
Check Your Skills 213
Chapter 3 Exercises 216
Chapter 4 Regression 231
The least-squares regression line 233
Facts about least-squares regression 243
Outliers and influential observations 248
Working with logarithm transformations 254
Cautions about correlation and regression 257
Association does not imply causation 263
Chapter 4 Summary 270
This Chapter in Context 272
Check Your Skills 273
Chapter 4 Exercises 276
Chapter 5 Two-Way Tables 289
Marginal distributions 291
Conditional distributions 295
Simpson’s paradox 303
Chapter 5 Summary 310
This Chapter in Context 311
Check Your Skills 312
Chapter 5 Exercises 315
Chapter 6 Samples and Observational Studies 325
Observation versus experiment 327
Sampling 332

Sampling designs 336
Sample surveys 345
Cohorts and case-control studies 352
Chapter 6 Summary 358
This Chapter in Context 360
Check Your Skills 361
Chapter 6 Exercises 364
Chapter 7 Designing Experiments 373
Designing experiments 375
Randomized comparative experiments 381
Common experimental designs 389
Cautions about experimentation 400
Ethics in experimentation 407
Discussion: The Tuskegee syphilis study 412
Chapter 7 Summary 419
This Chapter in Context 421
Check Your Skills 422
Chapter 7 Exercises 425
Chapter 8 Collecting and Exploring Data: Part I Review 438
Part I summary 442
Comprehensive review exercises 447
Large data set exercises 459
Online data sources 466
EESEE case studies 467
Part II: From Chance to Inference 468
Chapter 9 Essential Probability Rules 470
The idea of probability 472
Probability models 480
Probability rules 485
Discrete versus continuous probability models 490
Random variables 499
Risk and odds 504
Chapter 9 Summary 509
This Chapter in Context 511
Check Your Skills 512

Chapter 9 Exercises 515
Chapter 10 Independence and Conditional Probabilities 526
Relationships among several events 528
Conditional probability 536
General probability rules 541
Tree diagrams 553
Bayes’s theorem 560
Discussion: Making sense of conditional probabilities in diagnostic tests 568
Chapter 10 Summary 573
This Chapter in Context 575
Check Your Skills 576
Chapter 10 Exercises 579
Chapter 11 The Normal Distributions 589
Normal distributions 591
The 68–95–99.7 rule 597
The standard Normal distribution 603
Finding Normal probabilities 606
Finding percentiles 614
Using the standard Normal table 622
Normal quantile plots 629
Chapter 11 Summary 636
This Chapter in Context 638
Check Your Skills 639
Chapter 11 Exercises 642
Chapter 12 Discrete Probability Distributions 654
The binomial setting and binomial distributions 656
Binomial probabilities 660
Binomial mean and standard deviation 668
The Normal approximation to binomial distributions 672
The Poisson distributions 682
Poisson probabilities 685
Chapter 12 Summary 693
This Chapter in Context 695
Check Your Skills 696

- Chapter 12 Exercises 699

Chapter 13 Sampling Distributions 710
- Parameters and statistics 712
- Statistical estimation and sampling distributions 715
- The sampling distribution of x 728
- The central limit theorem 735
- The sampling distribution of p 753
- The law of large numbers 764
- Chapter 13 Summary 766
- This Chapter in Context 768
- Check Your Skills 770
- Chapter 13 Exercises 773

Chapter 14 Introduction to Inference 779
- Statistical estimation 782
- Margin of error and confidence level 793
- Confidence intervals for the mean μ 798
- Hypothesis testing 803
- P-value and statistical significance 809
- Tests for a population mean 820
- Tests from confidence intervals 827
- Chapter 14 Summary 829
- This Chapter in Context 831
- Check Your Skills 832
- Chapter 14 Exercises 835

Chapter 15 Inference in Practice 842
- Conditions for inference in practice 844
- How confidence intervals behave 851
- How hypothesis tests behave 858
- Discussion: The scientific approach 869
- Planning studies: selecting an appropriate sample size 873
- Chapter 15 Summary 888
- This Chapter in Context 890
- Check Your Skills 891
- Chapter 15 Exercises 894

Chapter 16 From Chance to Inference: Part II Review 901

Part II summary 904
Comprehensive review exercises 910
Advanced topics (optional material) 916
Online data sources 922
EESEE case studies 923
Part III: Statistical Inference 924
Chapter 17 Inference about a Population Mean 926
Conditions for inference 928
The t distributions 930
The one-sample t confidence interval 935
The one-sample t test 943
Matched pairs t procedures 952
Robustness of t procedures 959
Chapter 17 Summary 965
This Chapter in Context 967
Check Your Skills 968
Chapter 17 Exercises 971
Chapter 18 Comparing Two Means 980
Comparing two population means 982
Two-sample t procedures 989
Robustness again 1007
Avoid the pooled two-sample t procedures 1009
Avoid inference about standard deviations 1011
Chapter 18 Summary 1012
This Chapter in Context 1014
Check Your Skills 1015
Chapter 18 Exercises 1019
Chapter 19 Inference about a Population Proportion 1034
The sample proportion p 1037
Large-sample confidence intervals for a proportion 1040
Accurate confidence intervals for a proportion 1052
Choosing the sample size 1056
Hypothesis tests for a proportion 1060
Chapter 19 Summary 1069
This Chapter in Context 1071

Check Your Skills 1072
Chapter 19 Exercises 1075
Chapter 20 Comparing Two Proportions 1082
Two-sample problems: proportions 1084
The sampling distribution of a difference between proportions 1086
Large-sample confidence intervals for comparing proportions 1090
Accurate confidence intervals for comparing proportions 1096
Hypothesis tests for comparing proportions 1101
Relative risk and odds ratio 1109
Discussion: Assessing and understanding health risks 1118
Chapter 20 Summary 1123
This Chapter in Context 1126
Check Your Skills 1127
Chapter 20 Exercises 1130
Chapter 21 The Chi-Square Test for Goodness of Fit 1142
Hypotheses for goodness of fit 1144
Expected counts and chi-square statistic 1148
The chi-square test for goodness of fit 1152
Interpreting significant chi-square results 1159
Conditions for the chi-square test 1163
Chapter 21 Summary 1167
This Chapter in Context 1168
Check Your Skills 1169
Chapter 21 Exercises 1172
Chapter 22 The Chi-Square Test for Two-Way Tables 1183
Two-way tables 1185
Hypotheses for two-way tables of counts 1189
Expected counts and chi-square statistic 1192
The chi-square test 1200
Conditions for the chi-square test 1204
The chi-square test and the two-sample z test 1211
Chapter 22 Summary 1214
This Chapter in Context 1215

Check Your Skills 1216
Chapter 22 Exercises 1219
Chapter 23 Inference for Regression 1232
The regression parameters 1238
Testing the hypothesis of no linear relationship 1250
Testing lack of correlation 1255
Confidence intervals for the regression slope 1260
Inference about prediction 1263
Checking the conditions for inference 1270
Chapter 23 Summary 1278
This Chapter in Context 1280
Check Your Skills 1281
Chapter 23 Exercises 1284
Chapter 24 One-Way Analysis of Variance: Comparing Several Means 1298
Comparing several means 1300
The F statistic 1308
The analysis of variance F test 1312
Conditions for ANOVA 1322
Details of ANOVA calculations 1328
Chapter 24 Summary 1336
This Chapter in Context 1337
Check Your Skills 1338
Chapter 24 Exercises 1341
Chapter 25 Statistical Inference: Part III Review 1354
Essential concepts from Parts I and II 1356
Part III summary 1359
Comprehensive review exercises: inference selection 1366
Comprehensive review exercises: analysis and conclusion 1371
Large data set exercises 1381
Online data sources 1386
EESEE case studies 1387
Back Matter 1388
Notes and Data Sources 1388
Tables 1452

TABLE A: Random Digits 1452
TABLE B: Standard Normal Probabilities 1453
TABLE C: t Distribution Critical Values 1455
TABLE D: Chi-square Distribution Critical Values 1456
TABLE E: Critical Values of the Correlation r 1457
TABLE F: F Distribution Critical Values 1458
TABLE G: Critical values m* for Tukey pairwise multiple comparisons with 95% confidence level, k comparisons, and N – k degrees of freedom (df) 1462
Answers to Selected Exercises 1464
Some Studies Recurring Across Chapters 1513
Index 1516
A 1516
B 1517
C 1517
D 1519
E 1521
F 1522
G 1522
H 1522
I 1523
L 1524
M 1524
N 1525
O 1525
P 1525
Q 1527
R 1527
S 1528
T 1530
U 1531
V 1531
W 1531
X 1532
Z 1532

Symbols 1533
Part IV Optional Companion Chapters (available on the PSLS 4e Companion Website) 1535
Chapter 26 More about Analysis of Variance: Follow-up Tests and Two-Way ANOVA 1535
Beyond one-way ANOVA 1537
Follow-up analysis: Tukey pairwise multiple comparisons 1540
Follow-up analysis: contrasts 1549
Two-way ANOVA: conditions, main effects, and interaction 1556
Inference for two-way ANOVA 1566
Some details of two-way ANOVA 1579
Chapter 26 Summary 1586
Statistics in Summary 1587
This Chapter in Context 1589
Check Your Skills 1590
Chapter 26 Exercises 1593
Chapter 27 Nonparametric Tests 1602
Comparing two samples: the Wilcoxon rank sum test 1606
Matched pairs: the Wilcoxon signed rank test 1625
Comparing several samples: the Kruskal-Wallis test 1638
Chapter 27 Summary 1646
Statistics in Summary 1648
This Chapter in Context 1649
Check Your Skills 1650
Chapter 27 Exercises 1653
Chapter 28 Multiple and Logistic Regression 1662
Parallel regression lines 1664
Estimating parameters 1671
Using technology 1675
Conditions for inference 1682
Inference for multiple regression 1688
Interaction 1698
A case study for multiple regression 1708
Logistic regression 1719

Inference for logistic regression 1725
Chapter 28 Summary 1731
Statistics in Summary 1734
This Chapter in Context 1736
Check Your Skills 1737
Chapter 28 Exercises 1741
Back Cover 1760